

在线社交网络分析

Online Social Network Analysis

Binxing Fang Yan Jia Jin Xu Jianhua Li
Jiayin Qi Hongli Zhang Xindong Wu Bin Zhou 著
Shanlin Yang Changjun Hu Li Guo
Xueqi Cheng Xiangke Liao

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

This book focuses on the interaction among three elements of online social networks, i.e. structure, group and information.

The structure characteristics and its evolution mechanism, the formation and interaction of group behaviors, and the propagation models and evolution rules of information are discussed in details. This book provides an important theoretical foundation for social network analysis and research on network information dissemination.

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

图书在版编目 (CIP) 数据

在线社交网络分析 = Online Social Network Analysis: 英文/方滨兴等著. —北京: 电子工业出版社, 2017.9

ISBN 978-7-121-32745-2

I. ①在… II. ①方… III. ①互联网络—应用—人际关系学—研究—英文 IV. ①C912.11-39

中国版本图书馆 CIP 数据核字 (2017) 第 232105 号

责任编辑: 徐蔷薇 特约编辑: 马晓云

印 刷: 北京季蜂印刷有限公司

装 订: 北京季蜂印刷有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编: 100036

开 本: 787×1092 1/16 印张: 35.25 字数: 903 千字

版 次: 2017 年 9 月第 1 版

印 次: 2017 年 9 月第 1 次印刷

定 价: 98.00 元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888, 88258888。

质量投诉请发邮件至zlts@phei.com.cn, 盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式: xuqw@phei.com.cn。

PREFACE



In the 21st century, human beings are highly dependent on data, and have deeply fit into the information society, in which a huge function platform has been established by online social networks. Human beings are stating their viewpoints, making friends, and interacting on Twitter, Facebook, LinkedIn, Sina Microblog, WeChat and other social networks. Hundreds of million pieces of information are generated each day, and massive amounts of information is conveniently available to people. Online social activities are changing human behavior models and social formation, and social network data is becoming the most mature big data. By using the technique of big data, it is hopeful that people's understanding on user behaviors and social phenomenon, which are behind the big data of online social networks, may reach an unprecedented depth.

Online social networks analysis relates to computing science, sociology, management, psychology and many other subject areas. As a Chief Scientist of Project 973, i.e. "Fundamental Research of Social Network Analysis and Network Information Diffusion", and in my engagement in the research of online social networks, I deeply felt that this field lacked a treatise, which systematically elaborates the concepts, theories and techniques of online social network analysis from a multi-disciplinary angle. Hence, I organized team members of the Project 973, including National University of Defense Technology, Shanghai Jiao Tong University, Hefei University of Technology, Beijing University of Post & Telecommunications, Institute of Computing Technology, CAS, Peking University, University of Science and Technology Beijing, Institute of Information engineering, CAS, and Harbin Institute of Technology, the team compiled this book on the basis of the research findings of these teams and a systematic review of relevant theories and techniques at home and abroad, so as to provide theoretical, systematic and instrumental research guides for relevant researchers.

Starting from three core factors of the analysis of online social networks, i.e. "Structure and Evolution", "Groups and Interaction" and "Information and Diffusion", this book includes 12 chapters. Chapter 1 Introduction leads the entire book, and is written by myself. Chapter 2 to Chapter 4 are about the first core factor, namely, "Structure and Evolution": Chapter 2 Analysis and Modeling of Social Network Structure Characteristics

is written by Professor Jin Xu and Professor Hongli Zhang; Chapter 3 Detection Techniques and Approaches for Virtual Communities is written by Professor Jianhua Li; Chapter 4 The Evolution and Analysis of Virtual Communities is written by Researcher Xueqi Cheng. Chapter 5 to Chapter 8 are about the second core factor, namely, “Groups and Interaction”: Chapter 5 Analysis of User Behaviors is written by Academician Shanlin Yang; Chapter 6 Social Network Sentiment Analysis is written by Professor Bin Zhou and myself; Chapter 7 Individual Influence Analysis and Technique is written by Professor Yan Jia and myself; Chapter 8 Group Aggregation and Influence Mechanism is written by Professor Jiayin Qi. Chapter 9 to Chapter 12 are about the third core factor, namely, “Information and Diffusion”: Chapter 9 Information Retrieval for Social Networks is written by Professor Senior Engineer Li Guo; Chapter 10 Rules of Information Diffusion in Social Networks is written by Professor Changjun Hu; Chapter 11 Topic Discovery and Evolution is written by Professor Xindong Wu; Chapter 12 Influence Maximization Algorithms is written by Academician Xiangke Liao.

Sincere appreciation to the following experts and scholars participating in the data collection, content arrangement and achievement contribution of this book: Zhaoyun Ding, Xiaomeng Wang, Bin Wang, Yezheng Liu, Xiaodong Liu, Shenghong Li, Aiping Li, Lei Li, Shiyu Du, Peng Wu, Xiuzhen Chen, Wei Chen, Yang Yang, Lumin Zhang, Peng Shi, Yuanchun Jiang, and so on.

Thanks Associate Professor Shudong LI for the careful coordination and arrangement in the writing process of this book! Thanks Weihong HAN and Shuqiang YANG for working hard in the phase of reviewing and proofreading of this book!

CONTENTS



Chapter 1 Introduction	1
1.1 Social Network and Its Development	1
1.1.1 The Origin of Social Network	1
1.1.2 A Glimpse of the Development Procedure of Social Networks from the Perspective of Sociology	2
1.1.3 A Glimpse of the Development of Social Network from the Perspective of Anthropology	4
1.2 Development of Online Social Networks	5
1.2.1 Concept of Online Social Networks	5
1.2.2 Features of Online Social Networks	7
1.2.3 Development of Online Social Networks	8
1.2.4 Influences of Online Social Networks on People's Life	9
1.3 Background and Significance of Online Social Network Analysis	11
1.4 Scientific Questions of Online Social Network Analysis	13
1.4.1 Challenges of Online Social Network Analysis	13
1.4.2 Three Scientific Questions and Associated Researches	15
1.5 Organization of This Book	29
References	32
Chapter 2 Social Network Structure Analysis and Modeling	35
2.1 Introduction	35
2.2 Examples	36
2.3 Statistical Characteristics of Social Network	37
2.3.1 Degree Distribution	38
2.3.2 Average Path Length	39
2.3.3 Density	40
2.3.4 Clustering Coefficient	41
2.3.5 Betweenness	42
2.4 Social Networking Characteristics Analysis	43

2.4.1	Small-world Phenomenon	43
2.4.2	Scale-free Characteristic	47
2.4.3	Assortativity	53
2.4.4	Reciprocity	57
2.5	Social Network Structure Modeling and Generation	58
2.5.1	WS Model	59
2.5.2	Extension of WS Model	62
2.5.3	BA Model	63
2.5.4	Extension of BA Model	67
2.5.5	Other Models	70
2.6	Summary	74
	References	74
Chapter 3	Technologies and Approaches for Virtual Community Detection	78
3.1	Introduction	78
3.2	Theoretical Basis of Virtual Community Detection Technology	79
3.2.1	The Definition of Virtual Community	79
3.2.2	Development Process of Virtual Community Detection Algorithms	81
3.2.3	The Accuracy Indexes of Evaluation for Virtual Community Detection Algorithms	83
3.2.4	The Computational Complexity of Algorithms for Virtual Community Detection	88
3.2.5	Typical Data Sets Needed for Testing Virtual Community Detection Algorithms	89
3.3	Static Calculation Detection Algorithms for Virtual Communities	94
3.3.1	Modularity Optimization Algorithms	95
3.3.2	Multi-objective Optimization Algorithms	98
3.3.3	Algorithms Based on Probability Model	103
3.3.4	Information Coding Algorithms	107
3.4	Dynamic Calculation Detection Algorithms for Virtual Communities	112
3.4.1	Clique Percolation Algorithms	112
3.4.2	Agglomerative Algorithms Based on Similarity	116
3.4.3	Label Propagation Algorithms	120
3.4.4	Local Expansion Optimization Algorithms	125

3.5	Summary	128
	References	130
Chapter 4	Evolution Analysis of Virtual Communities	133
4.1	Introduction	133
4.2	Merging of Virtual Communities	134
4.2.1	Period Closure in Merging of Virtual Communities	134
4.2.2	Preference Connection in Merging of Virtual Communities	137
4.2.3	Aging Factors in Merging of Virtual Communities	142
4.3	Evolution of Virtual Communities	145
4.3.1	Accumulative Effect in Evolution of Virtual Communities	145
4.3.2	Structural Diversity in Evolution of Virtual Communities	149
4.3.3	Structural Balance in Evolution of Virtual Communities	154
4.4	Detection of Evolving Virtual Communities	156
4.4.1	Detection of Evolving Virtual Community Based on Direct Similarity Comparison at Adjacent Moments	156
4.4.2	Detection of Evolving Virtual Community Based on Evolution Clustering Analysis	158
4.4.3	Detection of Evolving Virtual Community Based on Laplacian Dynamics	159
4.4.4	Detection of Evolving Virtual Community Based on Clique Percolation Algorithm	161
4.4.5	Detection of Evolving Virtual Community Based on Trend Analysis on Node Behavior	162
4.5	Summary	163
	References	164
Chapter 5	Analysis of User Behavior	167
5.1	Introduction	167
5.2	Online Social Network User Adoption and Loyalty	168
5.2.1	Online Social Network User Adoption	168
5.2.2	Online Social Network User Loyalty	178
5.3	Individual Usage Behavior	189
5.3.1	General Usage Behavior	189
5.3.2	Behavior of Content Generation	195
5.3.3	Behavior of Content Consumption	206

5.4	Group Interaction Behavior	214
5.4.1	Relationship Selection of Group Interaction	214
5.4.2	Content Selection of Group Interaction	220
5.4.3	The Time Law of Group Interaction	222
5.5	Summary	226
	References	227
Chapter 6	Social Network Sentiment Analysis	233
6.1	Introduction	233
6.1.1	History of Sentiment Analysis	234
6.1.2	Sentiment Definition and Classification	235
6.1.3	Application of Sentiment Analysis	237
6.2	Sentiment Analysis Techniques	238
6.2.1	Semantic Rule-based Sentiment Analysis	238
6.2.2	Supervised Learning-based Sentiment Analysis	243
6.2.3	Topic Model-based Sentiment Analysis	249
6.3	Social Network Sentiment Analysis Techniques	251
6.3.1	The Sentiment Analysis Technique for Short Text	251
6.3.2	Sentiment Analysis Based on Collective Intelligence	255
6.3.3	Mining Techniques on Spam Opinions in Social Network	258
6.4	Extension and Transformation of Sentiment Analysis Technique	259
6.4.1	Sentiment Summary Technique	259
6.4.2	Sentiment Analysis Technology Based on the Mechanism of Transfer Learning	261
6.5	Summary	263
	References	264
Chapter 7	Influence Analysis and Its Technologies	267
7.1	Introduction	268
7.2	Influence Strength Calculation	270
7.2.1	Influence Strength Calculation Based on Network Structure	271
7.2.2	Behaviour-based Influence Strength Calculation	272
7.2.3	Topic-based Influence Strength Calculation	274
7.3	Identification of Influentials	277
7.3.1	Individual Influence Calculation Based Network Structure	277
7.3.2	PageRank	282

7.3.3	Individual Influence Calculation Based on Behavior	285
7.3.4	Individual Influence Calculation Based on Topics	289
7.4	Summary	291
	References	292
Chapter 8	Collective Aggregation and the Influence Mechanisms	294
8.1	Introduction	295
8.2	Mechanisms Engendering Collective Intelligence	297
8.2.1	Collective Intelligence	297
8.2.2	Self-determination Theory and Collective Intelligence	299
8.2.3	Conditions Engendering Collective Intelligence	301
8.2.4	Factors Influencing Group Intelligence	302
8.2.5	Analytical Models of Collective Intelligence	306
8.2.6	Simulation of Collective Intelligence in Social Networks	313
8.3	Mechanisms Engendering Group Polarization	323
8.3.1	Group Polarization	323
8.3.2	Social Comparison Theory and Group Polarization	325
8.3.3	Conditions Engendering Group Polarization	327
8.3.4	Factors That Influence the Formation of Group Polarization	328
8.3.5	Main Models of Group Polarization Analysis	331
8.3.6	Simulation of Group Polarization in Social Networks Without the Influence of Social Network Structure	342
8.3.7	Simulation of Group Polarization in Social Networks With the Influence of Social Network Structure	347
8.4	Summary	357
	References	359
Chapter 9	Information Retrieval in Social Networks	364
9.1	Introduction	365
9.2	Content Search in Social Network	368
9.2.1	Classical IR and Relevance Feedback Models	369
9.2.2	Query Representation in Microblog Search	379
9.2.3	Document Representation in Microblog Search	385
9.2.4	Microblog Retrieval Models	390
9.3	Content Classification	396
9.3.1	Feature Processing in Short Text Classification	397

9.3.2	Short Text Classification Algorithm	400
9.4	Social Network Recommendation	403
9.4.1	Brief Introduction to Social Recommendation	405
9.4.2	Memory Based Social Recommendation	407
9.4.3	Model Based Social Recommendation	413
9.5	Summary	421
	References	422
Chapter 10	The Rules of Information Diffusion in Social Networks	432
10.1	Introduction	432
10.2	Influencing Factors Related to Information Diffusion in Social Networks	434
10.2.1	Structure of Social Networks	434
10.2.2	Groups in Social Networks	435
10.2.3	Information	436
10.3	Diffusion Model Based on Network Structure	437
10.3.1	Linear Threshold Model	437
10.3.2	Independent Cascades Model	439
10.3.3	Related Extended Models	441
10.4	Diffusion Model Based on the States of Groups	442
10.4.1	Classical Epidemic Models	443
10.4.2	Infected Diffusion Models in Social Networks	445
10.4.3	Diffusion Models Based on Influence	447
10.5	Diffusion Model Based on Information Characteristics	448
10.5.1	Diffusion Analysis for Multiple Source Information	448
10.5.2	Competitive Diffusion of Information	450
10.6	Popularity Prediction Method	452
10.6.1	Prediction Models Based on Historical Popularity	453
10.6.2	Prediction Models Based on Network Structure	454
10.6.3	Prediction Models Based on User Behaviors	455
10.6.4	Prediction Models Based on Time Series	457
10.7	Information Source Location	467
10.7.1	Concept of Information Source Location	467
10.7.2	Source Location Methods Based on Centrality	469

10.7.3	Source Location Methods Based on Statistical Reasoning Framework	472
10.7.4	Multiple Information Source Location Methods	476
10.8	Summary	480
	References	481
Chapter 11	Topic Discovery and Evolution	485
11.1	Introduction	485
11.2	Models and Algorithms of Topic Discovery	487
11.2.1	Topic Model Based Topic Discovery	488
11.2.2	Vector Space Model Based Topic Discovery	502
11.2.3	Term Relationship Graph Based Topic Discovery	507
11.3	Models and Algorithms of Topic Evolution	512
11.3.1	Simple Topic Evolution	513
11.3.2	Topic Model Based Topic Evolution	515
11.3.3	Adjacent Time Slice Association Based Topic Evolution	518
11.4	Summary	519
	References	521
	Appendix	523
Chapter 12	Algorithms of Influence Maximization	527
12.1	Introduction	527
12.2	Basic Concepts and Theory Basis	528
12.3	Metrics of Influence Maximization	531
12.4	Classification of Influence Maximization Algorithms	533
12.5	Greedy Algorithm of Influence Maximization	533
12.5.1	Basic Concepts of Greedy Algorithm	533
12.5.2	Basic Greedy Algorithm	534
12.5.3	CELF Algorithm	536
12.5.4	Mix Greedy Algorithm	536
12.5.5	Other Greedy Algorithms	538
12.5.6	Summary of Greedy Algorithms	540
12.6	Heuristic Algorithms of Influence Maximization	540
12.6.1	Degree Discount Heuristic	540
12.6.2	PMIA Heuristic	542
12.6.3	LDAG Heuristic	542

12.6.4	Other Heuristics	543
12.6.5	Summary of Heuristic Algorithms	544
12.7	Extension and Deformation of Influence Maximization	544
12.7.1	Extension of Influence Maximization	545
12.7.2	Deformation of Influence Maximization	547
12.8	Summary	548
	References	549

Chapter 1

Introduction

1.1 Social Network and Its Development

1.1.1 The Origin of Social Network

Since the birth of human beings, they have been working together in farming and hunting, then a society is formed. With the development of society and deepening of communication, various relationships are established between people, and social relationships are developed to include friendship, production relationships, labor relationships, social interactions, etc. in addition to simple consanguinity and familial relationships. As social members interact with others in work, study, life, entertainment, and other activities, stable relationships are gradually formed and then a social network is generated. Just like what Mickenberg and Dugan said in 1995, “we all connect, like a net we cannot see”^[1].

In Wikipedia, the social network is defined as: “a social structure made up of a set of nodes. The nodes generally refer to individuals or organizations, and the social network stands for various social relationships. In the social network, relatively stable relationship systems are formed between members due to interactions, and the relationship systems may include friendships, classmate relationships, business partnerships, or race and faith relationships. By means of these relationships, the social network ties different people closely, from those meet each other occasionally to intimate family members and then to those in various social activities”^[2]. Since there are various social relationships in the

social network, the social graphical structure of social organizations or individuals tends to be very complex ^[2]. The complex relational structure affects interactions and associations between members, and thus has an effect on people's social behaviors.

From a historical perspective, the social network is the backbone to integrate people and the Internet. With the development of industrialization and urbanization and the rise of new communications technology, the society tends to be networked more and more closely. In 2012, Lee Rainie and Barry Wellman list the social network revolution, the mobile revolution, and the Internet revolution as three major resolutions affecting human society in the new period in their new book *Networked: The New Social Operating System* ^[3]. At present, the Internet, as an interactive platform playing an important role in mutual communication, interaction, and participation, has been developed significantly beyond ARPANET's original military and technical purposes. And the social network covers almost all forms of network services centered on human society, which allows the Internet to be developed from an application platform for research departments, schools, and governments/business into a tool for people to establish and develop relationships and to communicate with each other.

1.1.2 A Glimpse of the Development Procedure of Social Networks from the Perspective of Sociology

In 1842, French sociologist and positivism philosopher Auguste Comte (1798—1857) proposed a term^[4] “sociology”, defining two primary aspects of researches in sociology field: social statics and social dynamics. He is the first person who put forward studies on the society from mutual relationships between social actors. Auguste Comte considered that individuals are basic elements constituting the society, while individual properties in turn exert influence on the society's properties. Auguste Comte's contribution propels the development of sociology as a branch of science.

French sociologist Gustave Le Bon (1841—1931) claimed that the relationships^[5] among social members should be observed from a group perspective, and focused on the circulation of information among group members. Gustave Le Bon pointed out that when an individual becomes a member in a group, he or she will lose their identity as an individual. As a member in the group, people imitate others around them; as the group's ideas and behaviors get widely spread, individuals' ideas and behaviors are deeply influenced.

From the perspective of sociology, social network originates from “Sociology”

theory^[6] by German sociologist Georg Simmel (1858—1918). In the 1960s, with the beginning of the Cold War and social chaos pervasive in the western world, Georg Simmel's "Sociology" theory got rapid development in the west and became mature in the 1970s. Through the development process lasting for half a century, "Social Structure" theory^[7] in sociology has been widely applied in different fields including psychology, sociometric, sociology, anthropology, mathematics, statistics, and probabilism, and it is gradually formed into a set of systematic theories, methods, and technology, thereby becoming an important social structure study paradigm.

The popularity of the "social networks" concept exactly originates from his proper description of interaction of social relationships. Over the past century, sociologists have been using the metaphor of "social networks" to indicate various complicated social relationships. However, until 1950, the vocabulary began to be systematically used to indicate social communities having boundaries which are different from the traditional sense (such as villages and families) and a social category where people are regarded as separate individuals (such as gender and race). For example, the opinion of considering people in a café, colleagues working together, or people communicating with each other on the Internet as social communities having a boundary will lead to a wrong belief that they have a sense of belonging to their common group because they know each other. The truth is that people keep entering or exiting from a social network, and the social network has got a complicated structure.

In 1988, a well-known Canadian sociologist Barry Wellman proposed a relatively mature definition of social networks. Barry Wellman considered that the social network is a relatively stable system^[8] that is constituted by social relationships among certain individuals; that is to say, "network" is regarded as a series of social connections or social relationships linking the actors, and the relatively stable relationship mode constitutes the social structure. With continual expansion of the application scope, the concept of social networks has gone beyond personal relationships; network actors may be individuals, and may also be aggregation units, such as families, departments, and organizations.

The social network in the early stage mainly refers to off-line "social networks" established among individuals through acquaintanceship or working relationships, such as scientific research cooperation relationship networks, actor cooperation networks, and other relationship networks. Among them, the social relationship network of 34 members in a karate club in some university constructed by sociologist Wayne Zachary in the 1970s is a typical representation^[9] of early social networks.

With the development of Internet, the regional element reflected by the network structure is weakened; as a result, regional limitation in traditional off-line social networks becomes more and more weakened, and cross-regional online social relationships become an important pattern of social networks. After 2003, with the development of Web 2.0 technology, online social network media has attracted more and more attention from the people, who start to create accounts on online social media platforms, typically represented by Facebook, Twitter, Blog, and social networking sites, and add friends who they get acquainted with off line. Since then, off-line social networking begins to expand to network environments, and becomes an indispensable communication tool for people in their network life.

1.1.3 A Glimpse of the Development of Social Network from the Perspective of Anthropology

From the perspective of anthropology, the early study on the social network mainly included two social network modes: non-industrial society and industrial society.

First, in terms of the study on the social network in non-industrial society, the study on kinship of Lewis Henry Morgan (1818—1881), an American anthropologist, was the most representative^[10]. During studying Iroquois tribes, he found that kinship terms in Iroquois were totally different from those in modern America, while the kinship term system in other Indians was basically the same as that in Iroquois. He published a book *Systems of Consanguinity and Affinity of the Human Family* in 1871, pointing out that the kinship terms were not general, and different cultures had different kinship term systems. He solved the relation problem between the culture and kinship term system.

Alfred Radcliffe-Brown (1881—1955), an English anthropologist, inherited and developed the theory of Lewis Henry Morgan. He pointed out that the kinship system was a network of social relations, and was a composition of the network of total social relations. Such network of total social relations was called social structure. He advocated that a social network analysis method was used for analyzing the kinship relation to gradually form the structural functionalism theory^[11], leading the social network as a dominant concept in English anthropology.

But the traditional kinship study has two defects: first, it only focuses on individual members in the kinship relation and ignores the mutual relation among members; second, it focuses on the source and historical development of the kinship, and ignores the horizontal structure study. Lévi-Strauss (1908—2009), a French anthropologist, proposed a method of

studying the kinship relation from a structural perspective^[12], and summarized a binary opposition relation consisting of eight members in four groups of key relatives: husband and wife, brother and sister, father and son, and uncle and nephew. His method outlined a deep and general social network structure behind the kinship relation.

In addition to the study on the kinship relation, with colonialism collapsing and primitive society drifting away, the focus of the anthropology study shifted to wide agricultural society and social society. The study on the social relation also extended from the kinship relation to different social relations in urban, cities, enterprises, and organizations. In 1929, William Lloyd Worner (1847—1928), an American anthropologist, organized the “Yankee City” project to^[13] try to apply his method of studying Australian indigenous people to the study on American towns. He proposed a method of emphasizing the social class, individual interaction, and social network. His research had huge effects and a profound impact on later researchers.

Second, in the industrial society mode, the study of Max Gluckman (1911—1975), an important British anthropologist, was the most representative^[14]. During the study, he observed five factories. On one hand, he continued Radcliffe-Brown’s emphasis on the social structure. On the other hand, he started to focus on the wider social context where the workshop was in, and treated this as the key of the study. He found that informal organizations existed in the five factories. Although the relation mode between workers and managers were different, a social relation for spontaneous cooperation did not exist as the study said; instead, a large variety of conflicts existed between workers and managers.

In addition to being applied to the study on the modern industrial society, the social network was applied to the study on urbanization of third-world countries. Bruce Kapferer (1940-Present), an Australian anthropologist, studied the labor conflict in a mining company^[15]. Beside considering factors such as the interaction relation among workers, and the social network consisting of workers in his study, he analyzed the relation among the interaction of workers, the social network of workers, and key factory events such as strike action.

1.2 Development of Online Social Networks

1.2.1 Concept of Online Social Networks

With the rapid development of the Internet technology, people introduce the concept of

early social networking into the Internet, and create the online social networks for social networking services (SNS). The meaning of online social networks includes hardware, software, service, and application. Because a word group consisting of four words meets Chinese's word-formation habit, people customarily use the social network to replace the SNS.

The online social networks can be divided into four categories according to the research report^[16] on social computing of European Union:

(1) instant messaging applications, which are platforms for providing online real-time communication, such as MSN, QQ, AIM, Fetion and WeChat, and have mutual authentication and real-time push characteristics;

(2) online social applications, which are platforms for providing online social relationships, such as Facebook, Google+ (Google), RENN, Kaixin001 and Qzone, and have mutual authentication and non-realtime access characteristics;

(3) microblog-type applications, which are platforms for bi-directionally releasing short messages, such as Twitter, Sina weibo, Tencent weibo, and NetEase weibo and Sohu weibo, and have one-way authentication and realtime push characteristics;

(4) space-sharing type applications, which are Web 2.0 applications which can communicate with each other but are not tightly combined, such as forums, blogs, BBS, video sharing, social bookmark and online shopping, and have one-way authentication and non-realtime access characteristics.

The online social network is a social structure made up of a set of social actors and a set of ties between these actors in the information network. This social structure mainly includes three factors, including relationship structure, network groups and network information. The relationship structure of the social network is a network system formed via the connection of individual members of the society. Individuals are also referred to as nodes, and can be regarded as organizations, individuals, network ID, other entity or virtual individual with different meanings; however, the relationship between individuals can be family relations, movement behavior, send and receive messages and a variety of other relationships. Based on these relationships, the individuals in the social network self-organize a variety of virtual communities. The virtual community is a subset of social networks, and has a close connection between the nodes in the virtual community, and a sparse connection between the nodes of different virtual communities. On the basis of the relationships above, various information is transferred between the individuals, between an individual and a group, and between the groups in the social network. The constant

iteration process of this information transfer is information spreading in the social network. Due to the influences of the network structure and the information transfer, individuals cluster or get together in a certain virtual community for a certain event, affect, act and rely on each other, and purposefully act in a similar manner. This forms a group behavior of the social network.

The social network is a typical application of the Web 2.0 era, and is also a typical representative of sociality and initiative characteristics of the Web 2.0 era. The Internet of the Web 2.0 era is undergoing great change — from a series of websites to mature service platforms that provide network applications for final users^[17]. The contents in these platforms are produced due to the participation of each user; and personalized contents produced due to participation form current Web 2.0 world via people-to-people sharing. The social websites, such as Facebook and Twitter, are the masterpieces of the initiative participation of users in the Web 2.0 era. Facebook has been on the line for less than 8 years, and has owned more than 1,400 million users; and Twitter also has more than 500 million users. According to the report of each official website, until March 2013, the number of Sina Weibo users has exceeded 556 million.

Along with the incoming peak period of Web 2.0 applications, Web 3.0 also starts to emerge. Father of the Internet, Tim Berners-Lee, indicated^[18] that when the Internet develops into a semantic network covering a large number of data, people can access the incredible data resource, i.e., Web 3.0. In the network summit on 16 November 2010, Mary Meeker indicated that Web 3.0 is made up of “Social Networking, Mobile and Search”^[19]. In conclusion, the characteristics of Web 3.0 include: converting the Internet into a database, making a search engine intelligent, achieving the semantic network and the service-oriented architecture (SOA), and converting the Internet into a series of three-dimensional space (such as people, time and information). It can be said that who can lead Web 3.0, and who is the next actor of the network.

1.2.2 Features of Online Social Networks

Compared with traditional Web and information media applications, the online social networks mainly have the following new features.

(1) Immediacy: information can be released and received easily and quickly. A user can release and receive information over a phone or browser at any time from any place.

(2) Spreadability: “nuclear fission” type information dissemination. Once released, a

message will be immediately pushed to all followers by the system, and as long as being reposted, it will be disseminated to a new batch of followers at once. Thus, a “nuclear fission” type geometric spread trend is presented and opinion spread channels are created for ordinary people.

(3) Equality: everyone stands a chance to become an opinion leader. Compared with asymmetric information release and info receiver in traditional media, all users of social network services stand a chance to form opinion leadership over social networks, and thus play important roles in occurrence, development, spread, discussion stages of emergency events.

(4) Self-organization: they appear in the form of We Media and can quickly form virtual communities. That’s because individuals in social networks all have their own means and channels to provide and release information, and can quickly form online virtual communities by means of fast information dissemination of the social networks.

1.2.3 Development of Online Social Networks

In a certain sense, online social networks derive from people’s demands for social activities over networks. Figure1-1 illustrates the development history of online social networks. In 1838, Samuel F.B. Morse invented the Morse code, and then the telegram became a long-distance communication channel; this makes online real-time long-distance communication over social networks possible. In 1876, Bell invented a line switching mode that can be used in telecommunication networks, and thus the first practical telephone was officially produced. In 1969, ARPANET was invented in America. The first packet switched network in the world was then formally operated. In 1971, the first E-mail was sent by a researcher of Advanced Research Project Agency. In 1990, a researcher of European Organization for Nuclear Research developed a new World Wide Web (WWW) protocol, which marked the official birth of modern Internet. In 1994, Yahoo! GeoCities became the first Internet online community, where users can chat, make analysis and consultations. In 1999, Tencent QQ was launched. In 2002, Friendster was launched and became the first real online social network. In 2004, Facebook was launched. It now has developed to be the biggest online social network and already has 128 million active users by March 2014. In 2006, Twitter, the most popular microblogging website, was launched. It is one of top ten most visited Internet websites. Later, Sina Weibo and other domestic social networks were successively launched, which marked the further expansion and maturity of social networks.

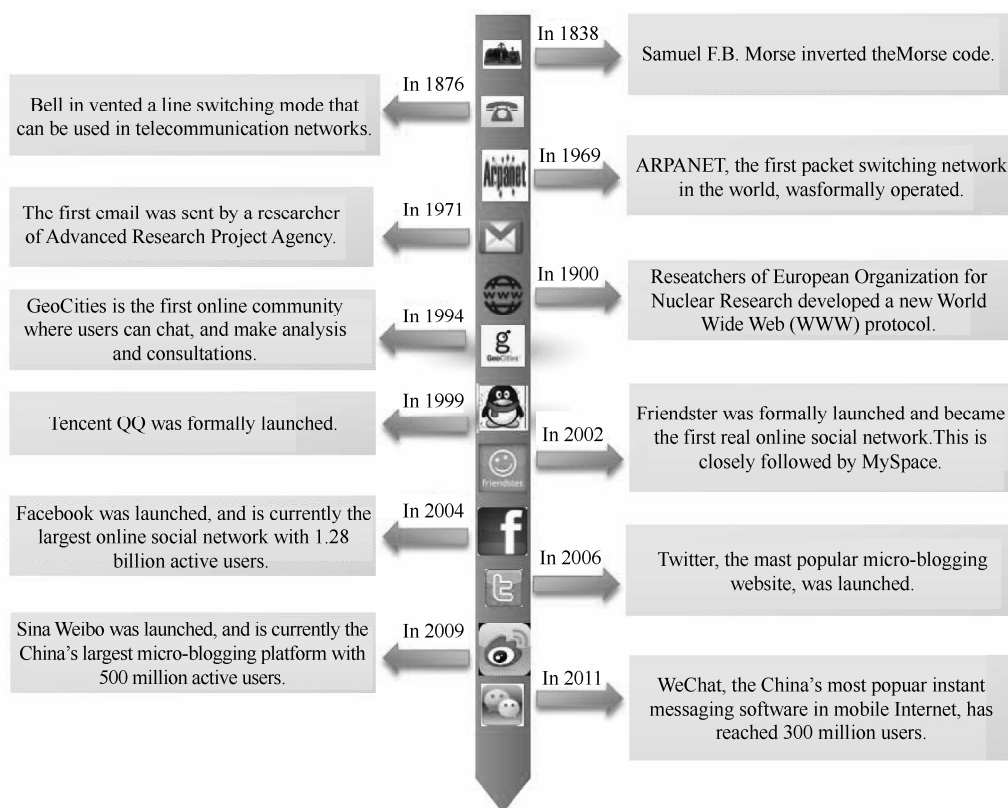


Figure 1-1 Development history of online social networks

Currently, social network applications are blooming. Based on a survey by Adobe, Facebook, which has been set up for more than 10 years, ranks first in top ten social networks by January 2014. It has about 1,400 million registered users and thus is called the third “populous nation”. Most of the registered users are from America, up to about 160 million, and what followed are users from Brazil, India, Indonesia, Mexico, Turkey, and Britain. Now, Facebook has 1,000 million monthly active mobile users. YouTube ranks second, having more than 1,000 million users. Chinese QQ Space and Sina Weibo, respectively having 623 million users and 556 million users, rank third and fourth. Following by are Twitter, Google+, LinkedIn, VKontakte (a Russian social network), Chinese RENN, WeChat in turn.

1.2.4 Influences of Online Social Networks on People's Life

The vigorous development of online social networks has not only greatly changed

people's way of life, but also brought negative effects to the society.

1. Online Social Networks Have Changed People's Way of life

Online social networks have gradually penetrated into all walks of life, affecting politics, education, economy, culture, and other aspects. In politics, microblogging has played a direct role in a number of government activities. In 2008, Obama had used Twitter for election activities, and his campaign team had moulded public opinion in microblogs to compete for votes. In education, more than 50 famous universities in the United States have released open classes in social networks to directly support distance education. Open classes are offered in Facebook and Twitter communities, and are integrated with educational resources such as MOOCs (Massive Open Online Courses). In economy, online shopping has become a mainstream way of shopping, and more than 70% of active adult users on social networks choose to shop online. Companies including Cole, Target and Ford have greatly improved their brand awareness by marketing on Facebook, and their turnover has increased by 10%. In culture, online social networks have changed people's way of life. Netizens can make friends, play games, and interact and cooperate with others without going out, and thus the so called "indoorsy life" is formed. In terms of social interaction and communication, social networks are sure to bring "social dividend" to conventional banks. The great potential of social networks allows domestic banks to develop strategies focusing on finding new profit growth points by using the social networks. Now in China, a WeChat Bank has been established under the cooperation of the China Merchants Bank with WeChat which is an emerging tool, to provide more convenient services to users with the help of WeChat, and this will also facilitate the application of WeChat to a broader field.

2. Online Social Networks Have Brought Negative Effects to the Society

Online social networks have also brought negative effects to the society while facilitating people's lives. In politics, some perpetrators deliberately create and spread rumors that are detrimental to national interests to affect social stability. For example, in 2011, the perpetrators incited the riots in London and other cities by using Facebook, and a "Rob Salt Tide" appeared in China because people were deluded by rumors on microblogging, which greatly affected the social stability. In education, evil forces educate young people to advocate violence and preach destructive individual heroism through social networks. In economy, perpetrators release false information through chat tools and

cheat customers through online shopping platforms. In culture, vulgar gangs disseminate online pornography and violent video content by means of social videos, instant messaging and other channels. In life, private information in social networks which make information transparent would be easily misused by undesirables, and as a result, people's normal life will be disturbed.

In terms of public opinion, event propagation on social networks contributes greatly to effects of public opinion. But adverse public opinion will have a huge impact on social stability. On February 15, 2011, riots broke out in Benghazi, Libya. Anti-government groups communicated with each other through Facebook and then set up a "liberal alliance", calling on people to join the anti-government forces, which had up to 15000 followers. On October 20, the Libyan leader Muammar Gaddafi was shot to death by the anti-government forces. Social networks played a role in fueling the early riots in Libya.

We can see from all given above that the essence of social networks is to help to form public opinion quickly, so as to influence people's thinking, affect people's world view, epistemic notions, values, and philosophy of life. In social networks, it is simple and convenient to release and receive information, every one has the right to speak online, various topics and views concerning national economy and the people's livelihood can be released at any time, and information can be spread in a way just like "nuclear fission" and may be overstated by opinion leaders to promote establishment of virtual communities for those having the same ideas and aspirations, and the masses can be organized and aroused quickly to participate in social activities, so a social mobilization force is formed.

1.3 Background and Significance of Online Social Network Analysis

Social network analysis relates to a calculable analysis method which integrates theories and methods of informatics, mathematics, sociology, management, psychology, and other sciences and is provided to understand formation of different social relationships, behavior characteristic analysis, and information dissemination laws. Social network analysis was first put forward by a well-known British anthropologist Radcliffe-Brown in his analysis and research on social structures. He appealed to carry out systematical research and analysis on social networks^[20]. With deepening analysis on social networks by sociologists, anthropologists, physicists, mathematicians, especially mathematicians in

graph theory, and statisticians, the theories, methods, and techniques formed in social network analysis have become an important social structure research paradigm. Thanks to online social networks' features of large scale, dynamism, anonymity, rich contents and data, etc., in recent years, analysis and research on emerging online social networks, such as social websites, blogs, and Weibo, have been flourishing, and play a decisive role in social structure research.

Social network analysis can function in many aspects, including political election, marketing, criminal chase, etc.

"Obama's victory in the United States presidential election of 2008" is a typical case of political election. On September 2008, Obama campaign created a social network analysis team to win the election. The team set up a donor database, an opinion poll database, and a network database. They grouped neutral electors based on their races, occupations, and other features, and then expressed different political propositions regarding different groups via emails, for example, they expressed propositions against racial discrimination for the black, expressed propositions on improving treatment for doctors, and expressed propositions on protecting workers' rights for building workers. These efforts produced a dramatical effect. On November 2008, according to data published by Gallup, the approval rate of Obama exceeds that of McCain by 11 percentage points, and Obama finally won that election. If it's just a trial in 2008, then Obama got familiar with the social network analysis in the election of 2012. Someone even call Obama as the first "social network president"^[21].

"Amazon recommendation system" is a typical case of marketing. Amazon is an American e-commerce company, on which people can shop, and books are its major products. The early Amazon recommendation system uses the approach of manual recommendation. But now, Amazon uses an automated recommendation system. The system first collects information of people on Amazon's website, including purchase history, browsing behavior, goods reviews, favorites contents of customers, then classifies customers with similar information into one group based on the information, and finally recommends goods to other users based on purchase of users from the same group. Application of the system improves Amazon's sales amount by 35%.

American LWAS information company, providing social media strategy consultation for the law enforcement agencies, says: "criminals will leave traces everywhere they went—in their cell phones, in their Twitter accounts, or in their Facebook". These traces can provide important clues for the law enforcement officers from both basic-level police departments and top government

agencies when they search for criminals on social media websites, such as Facebook, Myspace, Twitter, and YouTube^[22,23]. For example, New York Police Department set a Facebook special detachment to explore crime clues from social media^[24].

1.4 Scientific Questions of Online Social Network Analysis

1.4.1 Challenges of Online Social Network Analysis

An online social network has three central elements: first is its network structure, second is its group interaction, and finally the information dissemination. The three dimensions cover a wide range and involve several academic disciplines, such as informatics, mathematics, and management science. Therefore, the core of online social network analysis can be summarized as three elements: “Relationship structure”, “Network groups” and “Network information”. As illustrated in Figure 1-2, the three elements are correlated and interdependent: the “relationship structure” provides an underlying platform to network group interaction behaviors, and is the carrier of a social network; the “network groups” directly promote dissemination of network information and affect the relationship structure in turn, and are the subjects of the social network; and the “network information” and its dissemination are prerequisites of the social network, are also inducements and effects of group behaviors, can affect the change of the relationship structure, and are the objects of the social network.

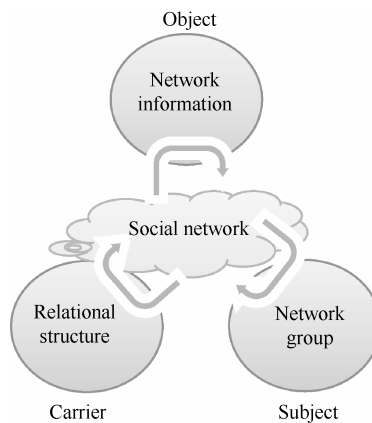


Figure 1-2 Relationship structure, network groups and network information in a social network

Since the study on the nature of large-scale social networks and the basic law of network information dissemination is still in a relatively primary stage, complete fundamental theories and methods on social networks and their information dissemination have not been proposed yet. For the three central elements, there are still a number of problems that are not clearly understood, and many challenges are waiting ahead.

Challenge I: the structure of an online social network is characterized in mass nodes, structural complexity, multi-dimensional evolution, etc. Specifically speaking,

- Mass nodes: Facebook has more than 1.28 billion active users all over the world, and is the “third most populous nation” ^[25].
- Structural complexity: there are various relationships among 536 million users of Sina Weibo, including concern/fans, mention, and forward.
- Multi-dimensional evolution: in a community of Twitter for discussing infectious diseases, evolving topics are complex and diverse within 30 days ^[26].

Challenge II: group interaction in online social networks has the characteristics of strong interactive evolution, public emotional upheaval, etc. The standpoints of the public are changing and their points of interest are evolving. Specifically speaking,

- Group clustering: a plurality of Twitter users published false information, raising the public panic event that a large number of migrant workers flee Mumbai, Bangalore, Chennai and other cities ^[27].
- Strong interactivity: in the community consisting of Sina Weibo users, the interactive relationships include mention, reply, concern/fans, and other direct relationships.
- Emotional variability: people are most positive at breakfast time of a day; after lunch, people get depressed gradually; and before bedtime, people’s emotions rise sharply ^[28].

Challenge III: information dissemination of online social networks has the characteristic of multisource concurrence, and the mutual influence makes the paths variable and content evolutionary. Specifically speaking,

- Multisource concurrence: for example, Kai-Fu Lee registered accounts on multiple social network platforms, and then published information on the multiple platforms concurrently.
- Mutual influence: KLM Royal Dutch Airlines used social media for marketing communication, and achieved good results through the mutual influence between Twitter users ^[29].

- Content evolution: in the Egyptian revolution, within an hour after the former President Hosni Mubarak announced that he would step down, the topic “Jan25” on Twitter spread and evolved so fast^[30].

1.4.2 Three Scientific Questions and Associated Researches

Regarding the aforementioned three challenges, from three new perspectives, i.e., “structure and evolution”, “group and interaction”, and “information and dissemination”, we have generalized in this book corresponding three scientific questions during online social network analysis, which are mainly reflected in three aspects: property and evolution mechanism of social network structure, formation and interaction law of social network group behaviors, and social network information dissemination law and evolution mechanism.

(1) Scientific question I: property and evolution mechanism of online social network structures.

How are the structures of social networks represented? What kind of representation method can reflect the essence of social networks while support computation and analysis at the same time? What kind of computation method can accurately depict the evolution of social network structures?

(2) Scientific question II: formation and interaction law of online social network community behaviors.

How can the existence and formation methods of groups be depicted in social networks? How can the interactive influence among groups be represented and measured? How should the influence exerted by the interactive process among groups on the evolution of communities be computed?

(3) Scientific question III: online social network information dissemination law and evolution mechanism.

How can the connotation of information be represented in a computable way? What are the methods for computing the dissemination process and status of information on social networks? How is the computation method used to depict mutual influence between information connotation and information dissemination?

Hereinafter, we will expound major issues and contents involved in the three scientific questions.

1. Structure and Evolution - Properties and Evolution Mechanisms of Social Network Structures

“Structure and evolution” corresponds to the first scientific question, i.e., properties and evolution mechanisms of social network structures. In this part, issues in three key aspects are to be solved, which are social network structure analysis and modeling, discovery of virtual communities in social network, as well as the evolution law of social networks.

1) Social network structure analysis and modeling

Social network structure analysis and modeling serve as the basis for all analysis. Social network structure analysis refers to analyzing many statistical properties through a statistical analysis method, including the distribution law of node degrees in a network, the intimacy degree of relationships, the intimacy degree of acquaintance relationship, the importance of a certain user for message delivery with all other users in the network. Social network modeling refers to adopting a structure modeling methods regarding the properties of social networks to study the mechanisms generating these properties, so as to deeply understand the inner law and essential characteristics of social networks.

The well-known Weak Ties theory studies social networks from the perspective of structure. Even famous examples in network structure studies are researches on small-world property^[31] in 1998 and discovery of scale-free property^[32] in 1999. The researches have greatly boosted the development in the network research field, and initiated a new trend in network researches. In recent years, the development of online social networks has driven people to search for its complicated inner mechanisms eagerly. Social network structures reflect relationships between individuals in social networks. Full understanding of social network structures provides support for other studies on scientific questions of social networks and lays the foundation for establishing basic theories and analysis methods for social networks. These important scientific questions include user impact analysis, studies on community interaction mechanism, information dissemination and evolution analysis, community discovery as well as studies on community formation mechanism. As the mobile Internet era arrives, the scale of users in online social networks grows with each passing day. Tens of millions of users as well as a growing number make it infeasible to represent social network structures through visualized methods. The ordinary research method is depicting network structures by virtue of network statistical characteristics; as a result, network structures can be learned through researches on

different network characteristics.

The significance of social network structure analysis and modeling lies on the one hand in generating, when it is difficult to get the real network, an analog network for replacing the real network to conduct researches, and the other hand in exploring the method and mechanism for generating a specific network structure. Therefore, social network structure analysis and social network structure modeling are two primary methods in social network structure studies, and are inseparable from each other. Network statistical property guide network structure modeling, and serve as a yardstick for the structure modeling method. In addition, the structure modeling process also discloses the inner mechanism of corresponding network statistical property. Therefore, the establishment of the social network structure model not only provides us with a replacement of the real social network, but also enables us to explore the generation mechanism of the real social network. The researches on social network structures serve as the foundation for social network researches. The continual development of this research field greatly propels the progress of social network science, or even the complex network science.

2) Virtual communities and discovery technology

Virtual community discovery is a must-have function for social network analysis. In the field of sociology, community refers to a personal relationship network^[33] formed by a group of people who engage in public discussions on the network and who have a deep relationship with each other after a period of time. The phenomenon that relationships are uneven exists in the social network; the relationships among some individuals are intimate while some relationships are estranged; as a result, a community form having closer relationships is formed by centering on a certain focus on a certain regular community, and the community form can be considered as a virtual community structure in social networks. Virtual community structure refers to a typical topological structure feature of the online social network. In online social networks, such as Sina Weibo and Facebook, the condition of close relationships among users can be found by digging the community, so as to obtain social relationships among users as well as social characters of the users; and by further combining analysis on users' viewpoints/behaviors and so on in the community, it is easier to understand network topological structure characteristics, disclose inner function characteristics of a complex system, and understand individual relationship/behavior in the community as well as the evolution trend.

Virtual community structures tend to disclose various connotations, such as modules, categories, groups, and teams. Various complex networks can be abstracted as graph

structures in mathematics; therefore, many scholars and researchers are trying to describe and depict the community structures by virtue of the mathematical tool, and provide their respective discovery methods^[34] from different perspectives. In essence, the discovery of the virtual community in online social networks is the process of dividing network nodes into several sub-graphs according to the close connection degree of their inner topological structures. During the early stage of discovering algorithm researches in the community, people defined some global target functions for measuring network community structure strength, and designed algorithms to optimize these functions to discover the community structure. The network data structure scale in the early stage was relatively small; understanding of the community structure is relatively onefold; and the requirement for time complexity of algorithms is not so high, either. With the development of online social networks, the processing technology of social network big data brings new opportunities and challenges for community discovery; and how to rapidly discover the virtual community structure and analyze hierarchy and overlapping of the community structure in heterogeneous and multi-dimensional social networks is a key research issue of group management control on a rapidly changing network era.

As a meso structure in social networks, the virtual community not only reflects high dynamism of the network on a micro level, but also reflects stability of the network on a macro level; this feature is a research basis for the formation and evolution mechanism of the virtual community in social networks. With the vigorous development of online social networks, analyzing large-scale network data to discover virtual communities present therein can facilitate understanding structural constitution of network virtual society and can also prevent and reduce many illegal behaviors that threat network security. In addition, the discovery of network virtual communities further exerts important influence on dynamical behaviors on the network. For example, the present of virtual community structures in online social networks exerts immense influence on message dissemination. Therefore, in the online social network environment, analyzing and discovering the virtual community structures present in the environment play an important role in understanding structural characteristics of network stratification organization, categorization of node individuals, as well as various network properties and dynamic behaviors.

3) Analysis on virtual community evolution

A virtual community has dynamical evolution property; and it is necessary to analyze and identify the evolution mechanism. Virtual community structures reflect local aggregation features of individual behaviors in the network; these virtual community

structures are not constant; because online social network structures keep evolving over time, virtual community structures also undergo continuous evolution. A large number of various explicit or implicit virtual community structures exist in online social networks, such as circles on renren.com and teams on douban.com, which keep evolving dynamically and continuously. The evolution of virtual communities is closely related to the functions of social networks in such aspects as diffusion, anti-disaster, cooperation, and synchronization, and also plays a fundamental role in the evolution of the social networks.

In fact, the change from static network analysis to researches on the evolution of dynamic networks is a new trend of social network researches in recent years. With the advent of the big data era, obtaining and analyzing large-scale dynamic network evolution data are made possible, which provides an important basis for evolution analysis of dynamic network virtual communities. In recent years, researches have focused on the evolution issues^[35] of social network dynamic virtual communities mainly from three aspects. The first aspect is the mechanism for the emergency of virtual communities or the reason for the formation of virtual communities; how is the emerging process of virtual communities restored or modeled according to these mechanisms. The next is which factors have exerted influence on the evolution process of virtual communities after the formation of them; because online social network structures and evolution process are complicated with many influencing factors, how to dig out the key factors during the evolution of virtual communities has become an important and challenging topic in researches. Existing researches mainly concentrate on the influence of three key factors on the evolution of virtual communities, i.e., accumulative effect of user individuals, structural diversity and structural equilibrium. After the formation reason of virtual communities and factors influencing the evolution are realized, the discovery algorithms of dynamic virtual communities can be researched then, i.e., how to identify the whole sequences of evolved virtual communities in dynamic networks.

The evolution of social network virtual communities is closely related to network functions, and closely affects dissemination of information on social networks and also plays a fundamental role in the evolution of social networks. As a result, researches on the evolution and analysis of virtual communities are of vital importance. Besides, the influence of other behaviors in online social networks on evolution dynamics of virtual communities, such as information dissemination, artificial control, emergency events, and other factors in online social networks, attracts more and more attention from people. As the big data era arrives, researches on these aspects will also enter a phase featured by rapid

development.

2. Group and Interaction: Group Behavior Formation and Interaction in Social Networks

“Group and interaction” corresponds to the second scientific question, namely, group behavior formation and interaction in social networks. This scientific question mainly deals with user behavior analysis, network sentiment analysis and individual influence analysis in social networks as well as key issues in collective aggregation and mechanisms that influence it. It mainly includes user behavior analysis, social network sentiment analysis and individual influence analysis as well as analysis on collective aggregation and mechanisms that influence it.

1) User behavior analysis

User individual behavior is the basic action in the community and needs to be modeled. User behavior in online social networks includes self-display, building relationships with strangers, sharing interests and information, releasing, searching, browsing and pushing information; interacting with different people based on various topics; and building interest community, learning and entertainment community, sharing knowledge, study and communication, and sharing happiness.

User behavior is an important research part of online social networks. User behavior in social networks is the willingness of users to make use of social network services on the basis of a comprehensive evaluation on their own needs, social impact and social network technologies, and a variety of use activities resulted therefrom. The research on user behavior in online social networks is mainly based on the following two ideas: using online social networks as a specific information technology, and studying users’ adoption and denial behavior and loyalty to the online social network technology; considering online social networks as a platform for providing various services and applications, and studying the characteristics and regularities that users use various services and applications. Considering online social networks as a specific information technology, researchers explore the effects of demographic variables, individuality traits, and emotional, cognitive and motivational factors as well as social, physical and technological environment on users’ adoption and loyalty to online social networks by using classical theories of behavioral study such as Technology Acceptance Model, Theory of Planned Behavior, Expectation Confirmation Theory, and Flow Experience Theory. Regarding online social networks as a platform for providing a variety of services and applications, researchers conduct studies on

individual usage behaviors such as self-presentation, micro-blog release, search, browse and comment, and user group interaction behaviors such as relationship establishment and content selection so as to reveal the underlying mechanism of content creation behavior and content consumption behavior in online social networks.

User behavior is the external performance of user motivation. Mastering the characteristics and regularities of user behavior in social networks to analyze the internal mechanism of user behavior in online social networks helps providers of online social network services to innovate social network service models, and helps to provide a theoretical basis for the monitoring and intervention of network public opinion.

2) Social network sentiment analysis

Sentiment analysis (aka opinion mining in this book) is a process of analysis, processing and induction of subjective information (positive, negative or neutral). Subjective information shows user's emotion or sentiment orientation. In social networks, everyone has different influences due to different emotional states.

Although many researchers realized the importance of sentiment analysis, techniques for sentiment analysis developed slowly before 1990s. One important reason is that the available data for analysis is very limited. With the rise of Internet, the Internet has become an important media for people to obtain information. Massive online news and reports are available for researchers to study sentiment analysis. Sentiment analysis has developed rapidly in the field of natural language processing. With the rise of Web 2.0 and social networks, users can express their views and opinions on online social media at any moment, which contribute massive corpus for sentiment analysis as well as bringing many new problems and challenges. Compared with news and reports, documents in social networks are short in length with irregular grammar. Besides, there are a large number of noise data as well as popular Internet slang. All these characters make sentiment analysis more difficult. At the same time, group characteristics in social networks and link and interaction among groups also bring a new research field to traditional sentiment analysis (the previous main resources for sentiment analysis are news and reports). Techniques of semantic-based sentiment analysis and supervised learning based sentiment analysis are gradually formed. At present, sentiment analysis has involved in many research areas such as natural language processing, short text mining and Web data mining, playing an important role in management science and sociology.

Sentiment analysis is an important part of social network analysis, and it is widely used in areas of product comments, public sentiment control and information prediction.

Sentiment analysis provides users with an emotional summary of historical evaluation, which enables users to quickly understand the product's evaluation information. In the public opinion monitor area, various topics and views related to national interest and people's livelihood can be released at any time. The interaction between virtual social networks and the real society increases direct impact on society, which directly affects national security and social stability. Analysis on people's emotions and attitudes in the network will play a very important role in maintaining national stability and promoting social development.

3) Individual influence analysis

Individual influence analysis is an important part of social network analysis. Individual social influence can be reflected through social activities among users, for example, the behavior and thought of users are changed under the influence of other people. In today's online social era, social networks have been a significant impact on people's daily life and behavior, a small number of malicious users and opinion leaders take advantage of social network services to make and disseminate public opinions. Opinion leaders make interaction between the media and Internet users for public opinion, and their views tend to affect a large number of fans and public opinion.

With the emergence of a large number of online social network services and user's participation, the relevant research on the individual influence analysis in social networks has attracted the attention of large number of scholars both at home and abroad. Since Katz and Lazarsfeld found that social influence plays an important role in social life and decision making in the 1950s^[36], influence analysis has been widely used in a number of areas, such as recommendation systems, social network information dissemination, link prediction, viral marketing, public health, expert discovery, incident detection and advertising etc.. Early work explored and analyzed the performance and related factors of influence in social activities, and made an in-depth study on the function model and generation mechanism of social influence. Many social phenomena related to influence and their underlying principles have been discovered. However, at that time, the sample space for the study was small, and the available data was limited. A large number of objective data was needed. With the popularity of online social networks, social network large data has brought new opportunities and challenges. How to find high influence users in heterogeneous, multi-attribute social networks, and to analyze the influence intensity among users in social networks, is a research focus for the information decision in the fast-changing network age. At present, in social networks, individual influence analysis

mainly includes the analysis of the influence intensity among users and the discovery of the influence individual. Therefore, how to find high influence users in heterogeneous, multi-attribute social networks, and to analyze the influence intensity among users in social networks, is a key problem.

Because network individuals in social networks gradually turn into the state that a few become opinion leaders and most follow the crowd, the study on the individual influence in social networks has important theoretical value and practical significance. The opinion leaders can take advantage of social network services to make and disseminate public opinions. The opinion leaders make interaction between the media and Internet users for public opinion, and their views tend to affect a large number of fans and public opinion. The opinion leaders play an important role in participation and guidance. In addition, influence analysis is widely used in a number of areas, such as recommendation systems, social network information dissemination, link prediction, viral marketing, public health, expert discovery, incident detection and advertising etc.

4) Analysis on collective aggregation and mechanisms that influence it

A group in online social networks is a virtual community. Collective aggregation is usually triggered by specific incentives. Many individuals in the real society gradually join and interact with each other to form a closely related community. A group shares information and interacts with other members through online social networks, and the individuals in the group can influence each other. Consequently, the Web 2.0 world is full of countless virtual groups with different sizes, various purposes, and dynamic variations. Although some online virtual groups are based on geographical distribution, most of them have no geographical restrictions, which is a distinct difference from the traditional concept of a group.

The study on groups has been a hot issue of sociology, psychology, economics, and management science. Since *The Crowd* (Gustave Le Bon) was published^[37] in 1897, scholars have sought an in-depth understanding of groups, especially group behavior. Every economic phenomenon, political decision or social transformation undoubtedly results from collective power. With the rapid development of the Internet, the network is increasingly showing a trend of integration between the real society and the virtual society. The biggest feature is that members of society become the subject of the Internet. At present, the collective aggregation and mechanisms that influence it in social networks mainly focus on collective aggregation mechanism, modeling method and evolution rules, followed by mining and constructing a group evolution behavior model through analysis on behavioral

and psychological motivation and influence factors of the social network group to reveal the inherent mechanism of social network group evolution. On the basis of these features, online social networks are more likely to engender some extreme collective aggregation behaviors: group intelligence and group polarization. Briefly, group intelligence is a group behavior in which many individuals generate problem-solving ability superior to their own through mechanisms such as competition and cooperation, differentiation and integration, and feedback and selection. Group polarization refers to the effect of discussions among group members on individual members' opinions or decisions in group decision-making situations, which leads to behavioral consistency within groups. In the past two years, Diggle (Tony Diggle, 2013) et al. showed how group intelligence can help in finding a solution to the global scarcity of water^[38]. David (David H. Zhu, 2013) et al. applied polarization to decision making in business or company seminars^[39]. They are both topics with great theoretical and practical values, which deserves further exploration.

The analysis on collective aggregation and mechanisms that influence it is of great significance for the monitoring and guidance to public emergencies. A great amount of network events show that after a public emergency occurs, when traditional media has not reported it yet or keeps silent, network media reports it in the first time and tracks the latest developments to quickly attract the audiences' attention. When Internet users browse the relevant information on the Internet, they will express their opinions and interact with others by means of follow-up comments or instant comments. A lot of manpower and time are required for organization and communication in reality; however, network group behavior can usually form a large scale in a short time.

3. Information and Communications-Information Communications Regularity in Social Network and Evolution Mechanism

“Information and communications” corresponds to the third scientific question: information communications regularity in social network and evolution mechanism. This scientific question mainly solves problems in terms of information retrieval in online social network, information communications regularity in social network, topic discovery and evolution, and influence maximization calculation method.

1) Information retrieval in online social network

Information retrieval refers to the process of finding materials (which are usually documents) meeting user information requirements from the set (which is usually saved in a computer) of massive unstructured data (which is usually in a form of text)^[40]. In addition

to search systems which are similar to us and representative of Google, the information retrieval system includes a classification system, recommender system, question answering system, etc. With the rapid popularization and development of the social network, information retrieval creates new resources and opportunities, and also faces new problems and challenges. The method of obtaining information for the emerging resource, social network, attracts broad attention of industry and academic field.

Facing information retrieval in online social network is an important study. Massive information is generated in social networks representative of Weibo every day. At present, a large variety of emergent events, news, and powerful topics have been reported via the social networks. The social network plays an important role in some great events (such as earthquake and aviation accident). It can quickly gather information from different organizations and propagate it to users. In addition, celebrity and friend effects in the social network give good support to information propagation. Even some products have good sales via the social network. Currently, related technology facing information retrieval in social network covers three aspects of content retrieval, classification, and social recommendation, is initially applied to different social networks, and produces user values and product commercial values. Compared with the traditional Web page, the social network document has limitations on text words, is required to be expressed in a particular way, and includes the social relation information of the author and among authors. These differences make it difficult for the traditional information retrieval technology to be directly applied to the social network. The current work is to explore and analyze authors, topics, and hyperlink information in the social network, and study the short text feature and communications mechanism in the social network, thus discovering many ways of improving content retrieval, classification, and social recommendation. But the social network is developing and changing, the present work is not fully practical. How to perfectly apply information such as the author, the relation network of authors, forwarding and replying, and hyperlink to the existing method is a key study. Therefore, there're still many problems and difficulties for retrieval of the social network. The study on retrieval provides academic values and application values.

There're many differences between information retrieval in online social network and traditional information retrieval. The features of the former one bring challenges and opportunities for the traditional information retrieval. The study on the information retrieval technology of the social network can not only provide better experience for users but also get the latest information so as to provide the theoretical basis for public decision

and public opinion guidance. Moreover, it can promote communications and propagation of political, economical, and cultural activities, and offer important social values and application values.

2) Information communications regularity in social network

Information communications is an activity for delivering, receiving, and feeding back information via symbols and signals, and is a process of exchanging opinions, ideas, and emotions by people to know and effect each other^[41]. Social network information communications specifies the information communications process made via the social network medium. Born by its flexibility and openness, the online social network gradually becomes an important center where information is propagated in modern society. The information communications in the social network becomes active to an unprecedented degree.

The rapid development of the social network provides rich data basis for researchers so that they have opportunities to study the information communications mechanism and know the information communications regularity on the basis of massive true data. Currently, the social network information communications mainly involves the network structure, community in network, and the propagated information. Related work is made on the basis of these factors. In terms of the information propagation model, the network structure-based research result mainly includes the independent cascade model, linear threshold model, and extension model; the community-based research result mainly includes the infectious propagation model and influence propagation model; and the information feature-based research result mainly includes the multi-source information propagation model, information competition propagation model, etc. In terms of popularity prediction, information such as the popularity trend, final popularity, and short popularity of the network content is predicted. The earliest method is mainly based on the historical popularity and user behaviors. In order to improve the prediction precision, researchers have proposed a propagation process and network structure-based method in recent years. In terms of information tracing, tracing is a technical means of widely collecting information in the social network and tracking the particular information to find the initial site or user who publishes the information, and the information propagation path. The representative method includes a centrality measurement-based tracing method and a statistical reasoning framework-based tracing method. In addition, for multi-source concurrency and incomplete observation, researchers have proposed a back propagation and node partition-based multi-source information tracing technology.

The study on the information communications regularity in social network can help us understand the social network and social phenomenon, and make us aware the topology structure, propagation capability, and dynamics behaviors of the complicated network. Moreover, it is helpful in model discovery, influential node identification, and personalized recommendation. The research result has wide applications to information recommendation of marketing and shopping sites, and public opinion monitoring and guidance. In terms of social benefits, social organizations and government organizations can publish information based on the information propagation features and regularity to improve management efficiency and transparency, and filter information to reasonably guide public. Therefore, the study on the information communications regularity in social network has theoretical values and application values.

3) Topic discovery and evolution

In the study on the social network, the topic refers to an influential event or activity, or all related events and activities. Events or activities refer to things happened in the particular time and place. News and events in each time moment, place, and language are reported in the social network, and propagated online without geographical boundary.

Topic discovery and evolution are important studies in mining of network texts. The TDT (Topic Detection and Tracking) project^[42] initiated by the Defense Advanced Research Projects Agency (DARPA) defined and illustrated the topic discovery and evolution comprehensively. The topic discovery and evolution refer to discovery of texts with the same topic in news at the very beginning, and generally use a clustering method. However, with the development of the social network, the traditional topic discovery and evolution can't adapt to the current network environment. We need to find a new method and way to complete such task. First, topic discovery and evolution in the social network do not only focus on news texts; instead, Blog, Weibo, etc. become research objects. Second, compared with the traditional text, the text in the social network has its distinctive features such as short Weibo, frequent use of Internet slangs, and informal words. These features make it urgent for us to use a new analysis method or improve the existing method to improve the analysis result. At present, the study on the topic discovery includes two types: one is to preset a topic to be monitored, and detect whether this topic appears in the social network. The other one is monitoring the new topic appearing in the social network. How to extract the topic which interests users from massive, dynamic, and multi-source social network data, recommend it, track the development and change in the topic, and get the event trend is a key study on the information decision in the rapidly changing network

times. Different from the topic discovery and evolution in the traditional media, as a new research topic, the topic discovery and evolution in the social network are not deeply studied and explored. The study on the topic discovery and evolution is still at early stages.

With the development of the social network, we shift our attention to the network. Network space becomes a new site where we publish messages, propagate message, and know information. Therefore, analyzing the network information, especially information in the social network to find the topic and events, knowing the occurrence and evolution of an online topic are valuable and meaningful for enterprises to carry out targeted network marketing, and the government to perform multi-granularity public opinion monitoring.

4) Influence maximization

Influence maximization is a key research focus in the field of information propagation in the social network, and has the purpose of discovering the set of nodes having the most information propagation influence so as to propagate the information in the social network to finally maximize the information propagation range. Influence maximization is widely applied to important scenes of our daily life, such as marketing, advertisement, public opinion pre-warning, water quality monitoring, Internet campaign, and emergency notification.

In recent years, influence maximization has been emphasized by academic field and industry at home and abroad. Relevant researches and discussions have been published on international top conferences such as SIGKDD (ACM Conference on Knowledge Discovery and Data Mining), WWW (International Conference of World Wide Web), AAAI (Association for the Advancement of Artificial Intelligence), and ICDM (International Conference on Data Mining). At present, the academic field has had a deep study on the influence propagation model. The independent cascade model and liner threshold value model are widely studied. The current social network has large scale, specifically, the number of nodes is large, the association relation among nodes is complex, and the dynamic nature of the social network becomes stronger. These features bring great challenges for solving the influence maximization problem. The running time, algorithm precision, and extensibility are important factors that need to be considered when solving the influence maximization problem in the current large-scale social network environment. The existing two mainstream calculation methods for influence maximization, greedy algorithm and heuristic algorithm, can't meet requirements of short running time and high algorithm precision simultaneously. Many greedy algorithms for influence maximization are proposed, such as BasicGreedy, CELF, MixGreedy, and CELF++, and have high solution precision. But they have long running time, and can't be applied to the current

social network that rapidly develops. Compared with the greedy algorithm, the proposed heuristic algorithms such as DegreeDiscount, PMIA, LDAG, and IRIE discover influential nodes quickly based on heuristic information, and significantly reduce the running time but still can't meet large-scale social network requirements. Moreover, the heuristic algorithm is far from the greedy algorithm in terms of precision so that the existing heuristic algorithm can't be applied to application scenes requiring high requirements on the running time and algorithm precision, such as marketing and water quality monitoring. Therefore, when the high precision is met, how to efficiently solve the influence maximization problem of the large-scale social network is an urgent and challenging study.

The study on the influence maximization of the social network can provide important economic benefits and social benefits. For example, for the social network-based word of mouth marketing and advertising, how to maximize brand promotion effects and propagation ranges by promoting commodities and advertisements to which users and propagating information and effects. For the water quality monitoring and epidemic monitoring, how to maximize the monitoring range and promptly discover water quality pollution and epidemic outbreaks by positioning which places for water quality monitoring and epidemic monitoring. In conclusion, the study on the influence maximization of the large-scale social network provides very important research and application values.

1.5 Organization of This Book

This book starts from three scientific questions in online social network analysis, and then profoundly and systematically elaborates the fundamental theories, key methods and technologies of online social network analysis at three levels “structure and evolution–group and interaction–information and dissemination” by answering the three questions. Chapters of this book are arranged as shown in Figure 1-3.

1. “Structure and Evolution”

Covering Chapters 2, 3, and 4.

Chapter 2 Social Network Structure Analysis and Modeling: it mainly introduces common network statistics characteristics of online social network analysis, summarizes and analyzes the general laws revealed from statistics characteristics of online social

network, such as small-world phenomenon and power law distribution, and focuses on introducing the structure modeling methods for social network.

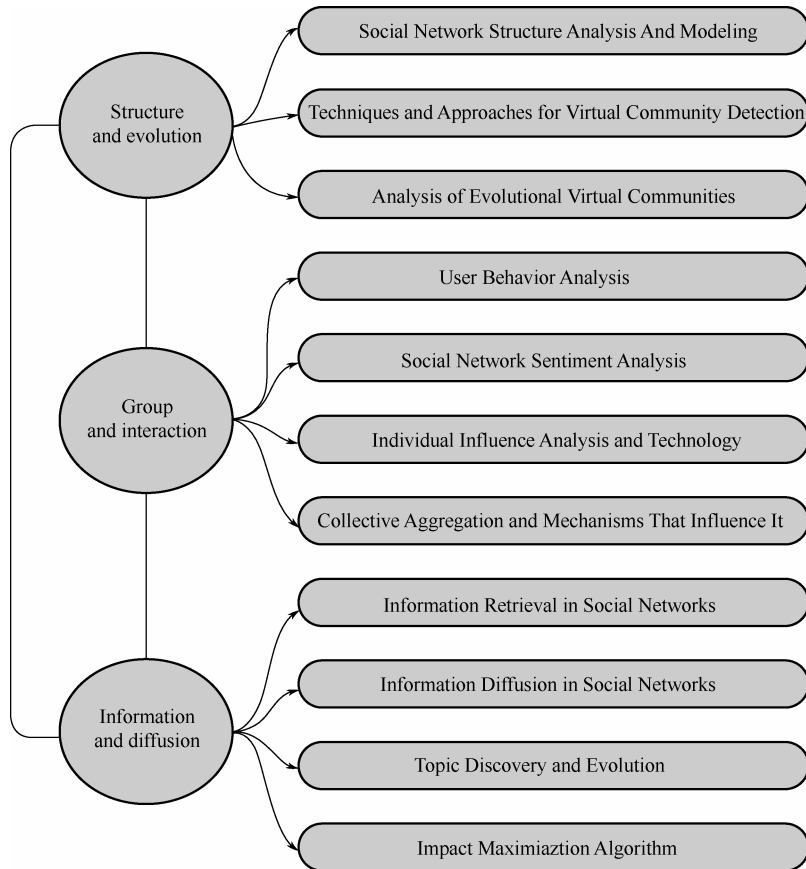


Figure 1-3 Correspondence between scientific questions and chapter contents of this book

Chapter 3 Techniques and Approaches for Virtual Community Detection: it summaries the definition of the virtual community structure, divides the currently prevailing virtual community detection algorithms into static computation detection algorithms and dynamic computation detection algorithms, and introduces related research works in virtual community detection algorithms on this basis.

Chapter 4 Evolution Analysis of Virtual Communities: it analyzes evolution of virtual communities mainly from three aspects: formation mechanisms of virtual communities, influencing factors on the evolution of virtual communities, and detection algorithms of virtual communities.

2. “Group and Interaction”

Covering Chapters 5, 6, 7, and 8.

Chapter 5 User Behavior Analysis: it introduces the influencing factors, modeling methods, and verification processes of social network adoption behavior and user loyalty research, the modeling methods of social network content creation behavior and content consumption behavior, as well as the analysis methods of group interaction selection behavior, content selection behavior, and interaction time laws.

Chapter 6 Social Network Sentiment Analysis: it introduces sentiment analysis problems emerging in social networks by combining structure characteristics and group interaction characteristics of social networks.

Chapter 7 Individual Influence Analysis and Techniques: it correspondingly explains social network individual influence research from three aspects: social network influence strength calculation, individual influence calculation, and influence test.

Chapter 8 Collective Aggregation and the Mechanisms that Influence It: it discusses collective aggregation and the mechanisms that influence it based on collective intelligence and group polarization.

3. “Information and Dissemination”

Covering Chapters 9, 10, 11, and 12.

Chapter 9 Information Retrieval in Social Networks: it introduces information retrieval research on social networks with three typical applications of search, classification, and recommendation, and mainly summarizes the current work on three aspects of query representation, document representation, and similarity calculation.

Chapter 10 Information Dissemination Laws in Social Networks: it elaborates network structure, group status, and information characteristic-based dissemination models and application examples, introduces a prediction method for dissemination statuses based on known information and its application examples, describes an information tracing method for tracking information dissemination sources, and analyzes some cases.

Chapter 11 Topic Discovery and Evolution: it detailedly introduces a topic model—the most important theoretical basis in research related to topic discovery and evolution, separately introduces the related scientific research statuses of two urgent and important research problems in topic discovery and evolution of social networks, and detailedly analyzes their respective technical characteristics.

Chapter 12 Influence Maximization Calculation Method: the modeling basis of influence maximization problems is a social network graph and its corresponding influence dissemination model. It summarizes influence maximization problems in the social networks and their main research methods, and focuses on research of greedy algorithm and heuristic algorithm for calculating influence maximization.

References

- [1] Linton C. Freeman. The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press, Vancouver, BC Canada, 2004.
- [2] Social Network[EB/OL]. Wikipedia [cited on March 12, 2014].
- [3] Lee Rainie, Barry Wellman. Networked: The New Social Operating System[M]. London: The MIT Press, 2012.
- [4] Ritzer, George; Goodman, Douglas. Sociological Theory (6/e)[M]. Taipei: McGraw-Hill, 2011: 32 and 33.
- [5] Gustave Le Bon (author), Dai Guangnian (translator). The Crowd (2nd Edition)[M]. Beijing: New World Press, 2011.
- [6] Georg Simmel. Sociology[EB/OL]. Wikipedia [cited on 28 February, 2014].
- [7] Jonathan H. Turner (author), Qiu Zeqi, Zhang maoyuan, et al. (translators). The Structure of Sociological Theory (7th Edition)[M]. Beijing: Huaxia Publishing House, 2006.
- [8] Barry Wellman, S. D Berkowitz. Social Structures: A Network Approach[M]. Cambridge University Press, 1988.
- [9] Zachary, W.W.. An Information Flow Model for Conflict and Fission in Small Groups[J]. Journal of Anthropological Research, 1977, 33: 452-473.
- [10] Lewis Henry Morgan. The Indian Journals, 1859-1862[M]. New York: Dover Publications, 1993.
- [11] A.R. Radcliffe-Brown (writer), Ding Guoyong (translator). Structure and Function In Primitive Society[M]. Beijing: China Social Sciences Press, 2009.
- [12] Claude·Lévi-Strauss (writer). The Elementary Structures of Kinship[M]. 1949.
- [13] John List (writer), Liu Jun (translator). Social Network Analysis (2nd Edition)[M]. Chongqing: Chongqing University Press, 2007.
- [14] Xia Xiyuan. Social Anthropology of Max Gluckman[D]. Beijing: Master's Thesis in Minzu University of China, 2010.
- [15] Bruce Kapferer. Strategy and Transaction in an African Factory[M]. Manchester: Manchester University Press, 1972.

- [16] Key Areas in the Public Sector Impact of Social Computing[EB/OL]. http://www.tno.nl/downloads/social_computing_impact_220609_final_report.pdf.
- [17] Web2.0. Wikipedia [cited on 22 June, 2014].
- [18] Victoria Shannon. A ‘More Revolutionary’ Web.: International Herald Tribune. 26 June, 2006.
- [19] Web3.0. Wikipedia [cited on 07 December, 2013].
- [20] A.R. Radcliffe-Brown. On Social Structure. The Journal of the Royal Anthropological Institute of Great Britain and Ireland, 1940, 70 (1): 1-12.
- [21] Liu Hong. Influence of Social Communication on Mass Communication[EB/OL]. http://qnjz.dzwww.com/tyzg/201401/t20140107_9480720.htm.
- [22] <http://v.ifeng.com/news/tech/201108/f46b19ae-a683-43a1-a22c-9a1a8610d690.shtml>.
- [23] <http://article.yeeyan.org/view/326883/280757>.
- [24] <http://news.ynxxb.com/content/2011-8/21/N95832747960.as>.
- [25] <https://alpha.app.net/hackernews/post/30997359>.
- [26] <http://www.nature.com/srep/2013/130828/srep02522/full/srep02522.html>.
- [27] http://www.chinadaily.com.cn/hqzx/2012-08/25/content_15705243.htm.
- [28] <http://blog.csdn.net/smarttony/article/details/6839076>.
- [29] <http://www.dmclick.com/daynews/detail.asp?id=1552>.
- [30] <http://globalvoicesonline.org/specialcoverage/2011-special-coverage/egypt-protests-2011/>.
- [31] Watts D J, Strogatz S H. Collective Dynamics of ‘Small-World’ Networks[J]. Nature, 1998, 393 (6684): 440-442.
- [32] Barabási A L, Albert R, Jeong H. Mean-Field Theory for Scale-Free Random Networks[J]. Physica A: Statistical Mechanics and its Applications, 1999, 272 (1): 173-187.
- [33] <http://zh.wikipedia.org/wiki/虚拟社群>.
- [34] Howard Rheingold (1993). The Virtual Community: Homesteading on the Electronic Frontier. London: MIT Press.
- [35] Easley D, Kleinberg J. Networks, Crowds, and Markets: Reasoning about a Highly Connected World[M]. Cambridge University Press, 2010.
- [36] Lazarsfeld P F, Katz E. Personal Influence: The Part Played by People in the Flow of Mass Communications[J]. Glencoe, Illinois, 1955.
- [37] Arthur F. Bentley. Review of The Crowd by G. Le Bon[J]. American Journal of Sociology. 1897 (2): 612-614.
- [38] Diggle T. Water: How Collective Intelligence Initiatives Can Address This Challenge[J]. Foresight, 2013, 15 (5): 342-353.
- [39] Goedert J D, Sekpe V D. Decision Support System–Enhanced Scheduling in Matrix Organizations

Using the Analytic Hierarchy Process[J]. Journal of Construction Engineering and Management, 2013, 139 (11).

[40] <http://www.cnblogs.com/AndyJee/p/3480273.html>.

[41] <http://zh.wikipedia.org/zh-cn>.

[42] <http://wiki.mbalib.com/wiki/TDT>.

Social Network Structure Analysis and Modeling

2.1 Introduction

Social network forms on the basis of social communication relationship between individuals in the society. Individuals, as nodes in a network, are those participants involved in social activities, which can be entities like organizations and persons or virtual individuals like network IDs, and the relationship between individuals can be family relations, behavior interactions, sending and receiving messages and so forth. The interactions between individuals include establishing or releasing acquaintance relationship, participating in the same topic discussion. These diverse individual behaviors in social network promote continuous evolution of network structure, which characterize social network with user group interaction, information dissemination and evolution, etc.

In respect of social network structure analysis, many studies have illustrated that a variety of real-world social networks have common structural characteristics of complex networks, such as small-world phenomenon, scale-free law, power-law distribution. In respect of social network structure modeling, many scholars have tried to conduct quantitative analysis on social networks using graph model, and many breakthroughs have been achieved in recent years, such as small-world model in which randomness is introduced into the regular network through reconnection mechanism, scale-free model featured by power-law distribution as a result of “preferential attachment” rule between nodes. Overall, the understanding for social network structure can be divided into the following three stages from shallower to the deeper: First, as there are usually tens of

millions of nodes and even more edges in social network, it is impossible to plot the real network structure, and statistical characteristics of network are usually needed, that is, when facing a large number of social network instances in real-world, we should obtain a preliminary description of the network structure through corresponding network parameters. Second, collecting and analyzing the static and evolution characteristics of certain network parameters in a large number of network instances to find out the general laws existing in the social network. Finally, establish corresponding network model to express these laws to understand the intrinsic mechanism in generating these laws.

This chapter is organized as follows: Section 2.2 firstly shows an example of a small online social network to facilitate the description and discussion. Following the three step-by-step stages introduced above, Section 2.3 mainly introduces the common characteristics of online social network, and define these characteristics and give corresponding formal descriptions. Based on Section 2.3, Section 2.4 summarizes and analyzes general laws revealed from statistical characteristics of online social network, such as small-world phenomenon, power-law distribution, etc. By introducing some important network structure models in complex network research and corresponding instances generated, Section 2.5 introduces the structure modeling methods for social network, including WS and its extension models, BA and its extension models as well as other models available for social network modeling, such as forest-fire model, Kronecker graph model and production model.

2.2 Examples

In social networks, there are many individuals participating in activities and the structure of network is complex. By graph theory, a mathematical tool, social network can be intuitively described as graph. Let $G=\{V,E\}$ denote the social network, where V denotes a node set with each node representing a person and E denotes an edge set with each directed or undirected edge set denoting the relationship between two persons. Figure 2-1 illustrates a typical structure of online social network with some nodes and edges in Sina Weibo.

In Figure 2-1, we select 8 users from the Top10 (ranked by the number of followers) from Sina Weibo to form a typical user following relationship network. We obtain many interesting conclusions by some simple analyses on Figure 2-1, such as these 8 persons have different backgrounds which can be divided into 5 classes: ① “Famous Big

V”(public intellectual) in scientific & academic circles and industrial circles Kaifu Li; ② Movie and TV stars Wei Zhao, Chen Yao, Kun Chen, Xinru Lin; ③ Famous writer Xiaoxian Zhang; ④ Actor, singer, racing driver Zhiying Lin; ⑤ Crosstalk master Degang Guo. It’s obvious that there are following relationships between 4 movie and TV play stars, thereby a tight social circle forms. Zhiying Lin and Xinru Lin from China TaiWan only follow each other. In Figure 2-1, Wei Zhao and Xinru Lin have the largest number of fans, and establish relationships with persons from different classes by following each other, which makes them the bridge linking persons in different area. Zhiying Lin and Degang Guo, Kun Chen and Chen Yao have unidirectional following relationship. These data were collected on 22:00 06 April 2014.

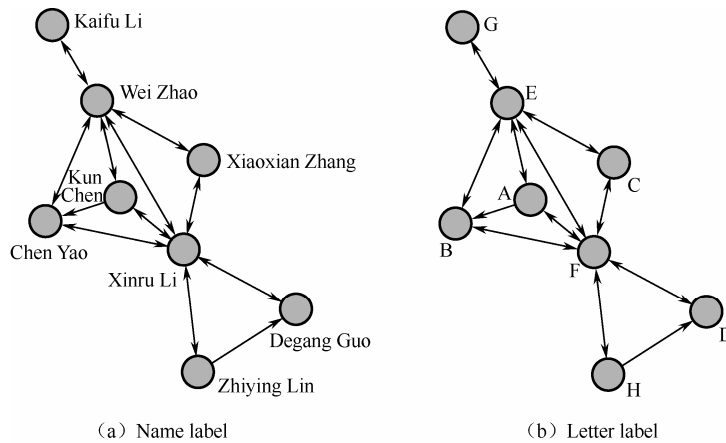


Figure 2-1 Social graph in Sina Weibo

We change names in Figure 2-1(a) into letters as shown in Table 2-1 for convenience, and obtain Figure 2-1(b).

Table 2-1 Comparison table for names and letters

Name label	Kun Chen	Chen Yao	Xiaoxian Zhang	Degang Guo	Wei Zhao	Xinru Lin	Kaifu Li	Zhiying Lin
Letter label	A	B	C	D	E	F	G	H

2.3 Statistical Characteristics of Social Network

In many math subjects, statistics is an important tool to describe uncertainty problems.

In the context of big data, analyzing and mining huge and various data demand the support of statistical knowledge. The data in social networks has typical characteristics of big data and often contains some noisy data, so statistical methods are needed for its description. In this section, we'll introduce some elementary statistical characteristics of social network, including degree distribution, average path length, clustering coefficient and so on, which will be the foundation of the following sections.

2.3.1 Degree Distribution

Degree is an important characteristic to describe the nature of node, which is defined as the number of edges connected to the node. In directed networks, degree is divided into out-degree and in-degree. For a node, the out-degree is the number of edges from it to other nodes and the in-degree is the number of edges from other nodes to it. The average degree of a network $\langle k \rangle$ is the average degree of all nodes in the network, which reflects the density of network, and degree distribution is used to describe importance of different nodes according to their distribution law of degree. By calculating the degree of each node and ranking nodes according to their serial numbers, we obtain the degree sequence of the network, based on which the degree distribution is obtained by calculating the frequency of node degree. It is worth noting that, though we lose the one-to-one correspondence between each node and its degree when calculating the degree distribution according to degree sequence, degree distribution can fully describe the law of degree distribution and identify different types of networks when the network scale is big. For example, ER random graph follows Poisson distribution, and complex network like online social network follow power-law distribution.

Degree distribution of nodes in networks can be described as follows:

(1) Function of degree distribution $P(k)$. $P(k)$ indicates the percentage of nodes with degree of k in the network.

(2) Function of cumulative degree distribution P_k . P_k indicates the probability distribution of nodes with degree no less than k , and its distribution relationship is

$$P_k = \sum_{x=k}^{\infty} P(x) \quad (2-1)$$

We call the network as a scale-free network following power-law distribution if its node degree follows distribution function $P(k) \propto k^{-\gamma}$ with power exponent of γ . The

cumulative degree distribution function P_k of the network follows power distribution $P_k \propto k^{-(\gamma-1)}$ with power exponent of $\gamma-1$.

We can obtain the node degree in the network by analyzing the letter labeling network in Figure 2-1(b), with degree sequence shown in Table 2-2 below. We can further obtain node degree distribution shown in Table 2-3 below by counting degree sequence.

Table 2-2 Degree sequence of nodes in networks shown in Figure 2-1(b)

Node	A	B	C	D	E	F	G	H
Degree	3	3	2	2	5	6	1	2

Table 2-3 Degree distribution of nodes in networks shown in Figure 2-1(b)

k	1	2	3	4	5	6	Total
$P(k)$	0.125	0.375	0.25	0	0.125	0.125	1

2.3.2 Average Path Length

The distance d_{ij} between any two users i and j in a social network is defined as the length of the path with the least edges between the users, also named as the shortest path length. As shown in Figure 2-1(b), for all the non-repeated paths between user G and user B, only G-E-B passes 2 edges, and the remaining passes more than 2 edges, so the shortest path length between G and B is G-E-B with length of 2.

Average path length L is defined as the average length of the shortest path between any two nodes in the network, it's also named as average length of the network or characteristic path length of the network. It describes the cost of information transfer between nodes in the network. In online social networks, L is always used to measure the relationship between users, and it represents the number of friends in the shortest path between any two users. It is calculated as follows:

$$L = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N d_{ij} \quad (2-2)$$

where N denotes the number of nodes in the network, d_{ij} denotes the shortest path length between nodes i and j .

Diameter D is defined as the maximum of all the shortest path lengths, i.e.

$$D = \max_{1 \leq i, j \leq N} d_{ij} \quad (2-3)$$

Effective diameter is often used instead of diameter in social networks. The reason is

that real social networks usually are not always fully connected since it includes many discrete nodes and connected branches. This issue can be avoided effectively when we select 90% of nodes, i.e. the minimum value which is larger than distances between at least 90% interconnected node pairs is called effective diameter [15]. The network diameter in Figure 2-1(b) is 3, with corresponding shortest path as G-E-F-H or G-E-F-D.

Example 2-1 Calculate the average path length and diameter of the network shown in Figure 2-1(b).

Solution: First, calculate the distances between all node pairs: $d_{GE}=1$; $d_{GB}=2$; $d_{GA}=2$; $d_{GC}=2$; $d_{GF}=2$; $d_{GH}=3$; $d_{GD}=3$; $d_{EB}=1$; $d_{EA}=1$; $d_{EC}=1$; $d_{EF}=1$; $d_{EH}=2$; $d_{ED}=2$; $d_{BA}=1$; $d_{BE}=1$; $d_{BC}=2$; $d_{BH}=2$; $d_{BD}=2$; $d_{AF}=1$; $d_{AC}=2$; $d_{AD}=2$; $d_{AH}=2$; $d_{CF}=1$; $d_{CH}=2$; $d_{CD}=2$; $d_{FH}=1$; $d_{ED}=1$; $d_{DH}=1$.

Then, substitute the data above into Formula (2-2) to calculate average path length

$$L = \frac{2}{8 \times (8-1)} \sum_{i \neq j} d_{ij} = 1.64$$

According to Formula (2-3), we obtain the diameter $D = \max_{1 \leq i, j \leq N} d_{ij} = d_{GH} = d_{GD} = 3$

2.3.3 Density

Density $d(G)$ describes the dense degree of interconnections between nodes in a network. It is the ratio between the actual and the maximal number of edges in the network, and is used to measure the dense degree and evolution trend of social relations in online social networks^[15]. The density of a network with N node(s) and L actual edge(s) is expressed as follows:

$$d(G) = \frac{2L}{N(N-1)} \quad (2-4)$$

The range of density is $[0,1]$. When the network is fully connected, $d(G)=1$. When there is no connected edge in the network, $d(G)=0$. There is nearly no network with density of 1, and the maximal density found in real network is $0.5^{[1]}$. Besides, the density of large-scale network is generally smaller than that of small-scale network. Densities of networks with different scale can't be compared directly, but they can be compared by absolute density formula^[2].

$$d(G) = M / [4SR^3 / 3D] \quad (2-5)$$

where D denotes the diameter, R denotes radius, S denotes perimeter calculated according to diameter.

In the network of Figure 2-1(b), we obtain density $d(G) = \frac{24}{8 \times (8-1)} = 0.43$ by substituting the number of nodes $N=8$ and real number of connected edges $L=12$ into Formula (2-4).

2.3.4 Clustering Coefficient

Clustering coefficient is used to describe the degree that nodes connected to the same node in a network are also adjacent nodes. For a node v_i , its clustering coefficient C_i describes the average probability that it connects to adjacent nodes. k_i denotes the number of neighbors connected to v_i and e_i denotes the actual undirected edges existing among k_i neighbors. It's easy to know that $\frac{k_i(k_i-1)}{2}$ is the maximum of links among these k_i neighbors, thus clustering coefficient of node v_i can be described as follows:

$$C_i = \frac{2e_i}{k_i(k_i-1)} \quad (2-6)$$

Clustering coefficient of a node has intuitive meaning in social networks. Informally, it's the probability that friends of the same person are also friends. It reflects the closeness of acquaintance among a person's friends in his/her circle of friends. Since most persons in one's circle of friends are their classmates, colleagues and relatives, it's highly possible that they know each other. Thus, social networks have strong aggregation. Average clustering coefficient is used to describe the aggregation of network.

Average clustering coefficient is the average of the clustering coefficient of all nodes in a network, and the formula is as follows:

$$C = \frac{1}{|V|} \sum_{i \in V} C_i \quad (2-7)$$

where $|V|$ denotes the number of nodes in the network, C_i denotes the clustering coefficient of v_i with range of value as $[0,1]$, $C=0$ indicates that there is no edge in the network; $C=1$ indicates that the network is fully connected.

The average clustering coefficient describes the probability that each two users of any three users know each other, which reflects the closeness of acquaintance among users in

the network.

Example 2-2 Calculate clustering coefficient of every node in and the average clustering coefficient of the network shown in Figure 2-1(b).

Solution: In Figure 2-1(b), nodes B, E and F directly connect to node A, the maximal number of probable edges among the three nodes is $3 \times (3 - 2) = 3$, and there are 3 edges in fact, thus the clustering coefficient of node A is $C_A = 3 / 3 = 1$ according to Formula (2-6). Similarly, the clustering coefficients of other nodes are shown in Table 2-4.

Table 2-4 Clustering coefficient distribution of nodes in the network shown in Figure 2-1(b)

Node	B	C	D	E	F	G	H
C_i	1	1	1	1/6	1/6	0	1

According to the clustering coefficient of each node shown in Table 2-4, we obtain clustering coefficient of the network shown in Figure 2-1(b) by Formula (2-7).

2.3.5 Betweenness

Betweenness describes the capacity of nodes to be part of the shortest paths in the network. Betweenness of a node (or an edge) is the sum of probabilities that shortest paths go through the node (or the edge), and reflects the impact and centrality of a node in a network. Assume that the number of the shortest paths between nodes i and j is δ_{ij} , and the number of the shortest paths going through a node k is $\delta_{ij}(k)$, then the ratio $\delta_{ij}(k) / \delta_{ij}$ is able to describe the importance of node k between i and j . On this basis, the betweenness of node k is defined as follows:

$$C_B(k) = \sum_{i \in V'} \sum_{j \neq i \in V'} \frac{\delta_{ij}(k)}{\delta_{ij}} \quad (2-8)$$

Betweenness is used to evaluate the capacity of traffic overhead in the Internet. Bigger betweenness C_B of a node indicates that more information can go through it during information dissemination. Some hub nodes used for data transmission usually has high usage rate with bigger data volume, which cause network congestion. In social networks, degree is used to measure the importance of a node. Though the degrees of nodes connecting different communities may be small, but they are extremely important, thus their betweennesses are usually relatively big. Hence, betweenness is usually used to evaluate the importance of a user in the information transmission between all user pairs in

social networks.

Example 2-3 Calculate betweenness of each node in the network shown in Figure 2-1(b).

Solution: According to the definition of betweenness, in the example shown in Figure 2-1(b), only nodes F and E are on the shortest paths between other node pairs in the whole network. The number of shortest paths going through node F is 10, i.e. A-F-D, B-F-D, E-F-D, C-F-D, G-F-D, A-F-H, B-F-H, E-F-H, C-F-H and G-F-H. Besides, there are two shortest paths (from node B to C and from node A to C) going through node E or F, so we obtain $C_B(F) = 8 + 1/2 + 1/2 = 9$ by formula (2-8). The shortest paths going through E are H-E-B, H-E-C, H-E-A, H-E-F, H-E-G, H-E-D, B-E-C, A-E-C, so the betweenness of node E is $C_B(E) = 6 + 1/2 + 1/2 = 7$. The betweennesses of all nodes are shown in Table 2-5.

Table 2-5 Betweenness distribution in the network shown in Figure 2-1(b)

Node	A	B	C	D	E	F	G	H
C_B	0	0	0	0	7	9	0	0

2.4 Social Networking Characteristics Analysis

This section describes and analyzes the regularities represented by the statistic characteristic in the social network. We focus on two important statistic characteristic in the topological research on social network: small-world phenomenon and scale-free characteristic. Small-world phenomenon indicates that the distance among persons in the social network are very short, which is proved by calculating the average path length in the network. Scale-free characteristic indicates that the node degree in the social network follows power-law distribution. Then, this section introduces assortativity and reciprocity of the social network. In addition, we also introduce several typical online social networks, and analyze the performance of these characteristics in different online social networks.

2.4.1 Small-world Phenomenon

Two people with long geographical distance tend to have shorter social relation interval. Sometimes people may find that someone seems “distant” is actually “very close” to you. In 1929, the Hungarian author Frigyes Karinthy put forward six degrees of segmentation for the first time in his short story “The Chain”. He said, although there are

great physical distances between individuals around the world, the increasingly stronger human relations makes the actual social distance between each other much more smaller. Two strangers can establish connection through five persons at most. In 1967, Stanley Milgram, a social psychology professor at Harvard University, summarized and put forward the famous “Six Degrees of Separation Hypothesis” by designing a letter delivery experiment^[5]. Milgram randomly chose 296 volunteers in Omaha, Nebraska as the initial sender to mail the letter and asked them to give this letter to a stockbroker in Boston. In the experience, each person made contact alone. Milgram told each sender the recipient’s information, including name, location, occupation, and if they didn’t know the recipient, they send the letter to an acquaintance that may know the recipient. In this way, the chain of the sender forms, and each member of the chain was trying to send this letter to their friends, family members, colleagues or acquaintances to transfer the letters to the recipient as soon as possible. Professor Milgram found that 60 chains eventually reached the recipient, with average steps of 6. Milgram drew the conclusion: any two persons can obtain in touch through an average of 5 acquaintances within 6 steps. The study unprecedentedly proved that human society is a small-world network with a shorter path length characteristic. Like the classic lines in John Guare’s movie “Six Degrees of Segmentation” shot after more than 20 years: “I am able to obtain in touch the President of the United States or a boatman in Venice as long as the right five persons between us are found”.

Small-world phenomenon was further confirmed by two famous real experiments^[39], i.e. interesting Kevin Bacon game and Erdős number. In 1997, 3 American students invented the Kevin Bacon game, and they considered Kevin Bacon, a movie actor, as the center of the filmdom. In Kevin Bacon game, if a movie costars a person and Kevin Bacon, his/her Bacon number is 1. The majority of actors/actresses in the world can build a direct or indirect acquaintance relationship with Bacon within six steps, i.e. the Bacon number less than or equal to 6. For example, the Bacon number of Ziyi Zhang, a famous actress in China, is 2 as she and Laurence Fishburne both participated in the animation “Teenage Mutant Ninja Turtles” dubbing, and Laurence Fishburne cooperates with Kevin Bacon in “Mystic River”.

In fact, after analyzing more than 1.7 million actors/actresses, only 260 people’s Bacon number is greater than 6, and the average Bacon number of all actors/actresses is 3. Paul Erdős, a famous mathematician, has put forward the random graph theory. Erdős numbers among mathematicians mean: Erdős himself marked as 0, people who had directly

worked with Erdős marked as 1, and persons marked as 2 if they directly worked with the persons with Erdős number of 1. If a person has multiple Erdős numbers, the smallest shall be chosen and so on. However, no matter how far the distance of industry and directions between other mathematicians and Erdős, their Erdős numbers are amazingly small. For example, Einstein's Erdős number is 2, Fermi's Erdős number is 3, Pauli's Erdős number is 4, and Heisenberg's Erdős number is 4. Erdős number of Bill Gates who only published one article on information theory is 4. Besides, social networks tend to show high clustering characteristics. For example, students of the same major in the school may know each other, and student of the same class may be friends. "Triple Transfer Ratio", the popular concept in sociology, describes a common phenomenon that a person's friends may also be friends.

That is to say, networks with social relationship tend to have relatively large aggregation, just like Milgram's experiment on small-world phenomenon. In 1998, two young physicists, Duncan Watts and Steven Strogatz^[24], published a landmark article in the "Nature" in which they put forward the concept of small-world network and set up a small-world model.

Take Figure 2-1(a) as an example, most shortest path lengths between stars are 1 or 2. The largest length of the shortest path is Degang Guo and Kaifu Li, with distance in Sina Weibo of $d(G, D)=3$. According to Example 2-1 listed in the length of the shortest path between all nodes, we can obtain the shortest path length distribution of the network in Figure 2-1(b) are shown in Table 2-6 below. $P(d)$ denotes the probability that the shortest path length equals to d :

$$L = \sum_{0 < i < N} dP(d) = 1 \times \frac{3}{7} + 2 \times \frac{1}{2} + 3 \times \frac{1}{14} = 1.64$$

Table 2-6 Distribution of star network in Sina Weibo

Shortest path length d	1	2	3
$P(d)$	3/7	1/2	1/14

We can see that the distance between any two stars in the star network in Sina Weibo is less than 2.

As the extension of the social network, the online social network also features significant small-world phenomenon. Shorter average path length of the network results in more evident small-world phenomenon. Compared with the traditional complex networks, the online social network has shorter average path length and effective diameter, with its effective diameter far less than the Web^[3] and average path length only 1/3 of the Web.

Yong-Yeol Ahn et al.^[4,23] studied the average path length distribution of the Cyworld, an online social network. We use the entire network topology in December 2005 of Cyworld, the biggest social network site for making friends in South Korea, as our data set provided by operators, which is a complete data set containing 12 million user nodes and 190 million edges composed of friend relationship. The measurement of the average path length of the network is divided into two steps. First, sample different number of seeds to randomly obtain 100, 2000 and 3000 nodes, then compare their data with the original data containing all nodes. Second, use BFS to obtain the average path length of users as shown in Figure 2-2. We can see that the average path length of sampling network converges to the average path length of entire network, and the shortest path length between most nodes is 4~5, while the average path length between the seed node and more than 90% nodes is less than 6. When measuring the average path length of the large-scale user graph, sampling is usually used to reduce the time for BFS. Jure Leskovec and Eric Horvitz^[38] screened the Microsoft MSN chatting records in a month of 2006, and analyzed 30 billion pieces communication information of 240 million active users. As the data was too large, they random chose 1000 users by sampling and calculated the average path length of the network. They found that 48% users can be associated within 6 steps, and the average path length of the MSN network is only 6.6.

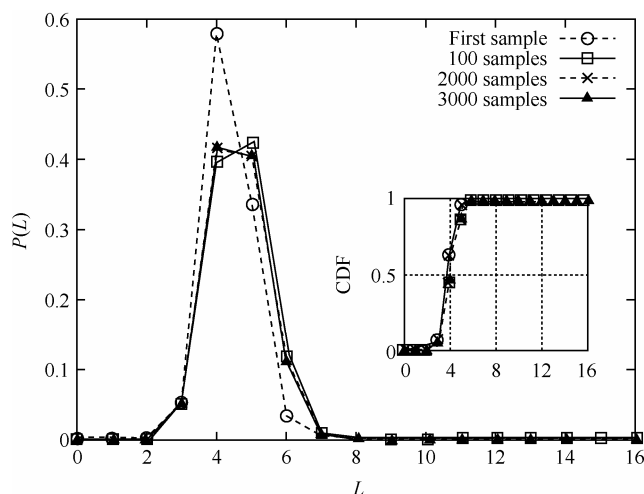


Figure 2-2 The distribution situation of the average path length of the Cyworld, an online social network in South Korea^[4]

There are some differences between the average path length of different types of online social networks^[3,4]. Those mainly for sharing has larger length than those mainly for

making friends, which online social networks mainly for making friends usual has smaller average path length and closer social ties, as well as more significant small-world phenomenon. With the development of online social networks, online world seems to be closer than “Six Degrees of Separation” as the online social network further narrows the distance between persons. In 2011, data team of Facebook counted the entire network dataset (approximately 720 million user nodes, 69 billion connected edges of friend relationship) found that the average path length between any two users is only 4.74^[6]. In the online social networks like microblog, users can establish connected edge by following other users. According to research of Sysomos, a social media monitoring company, on 5.2 billion similar connected edge relationship, the average path length of Twitter is 4.67. On average, about 50% persons are only four steps away on Twitter, and any individuals from two networks may associate within five steps.

2.4.2 Scale-free Characteristic

In nature and social life, events interesting scientists often have a typical scale, and changes of individual scale in the vicinity of the characteristic scale are very small. For example, a person’s height distribution, the number of passengers waiting in the bus stand and so on follows Poisson distribution. As shown in the Figure 2-3(c), most individual data concentrate in the vicinity of the average degree of the network $\langle k \rangle$ and tend to node with the number of individuals far away from average value decreasing exponentially, thus we call this $\langle k \rangle$ as the characteristic scale of network with homogeneity. Degree distribution of some real networks is quite different, such as the degree distribution of wealth, national population and number of friends on dating site. Individuals are quite different from each other in data systems of these real networks. Most nodes have a small amount of connected edges while few nodes have a large number of connected edges. The network present heterogeneity due to the lack of a unified measurement scale, and for those characteristic scales without limited measurement distribution range in node degree distribution, we call this feature as scale-free. In the random network as shown in Figure 2-3(a), the average node degree is about 2. However, scale-free network as shown in Figure 2-3(b), some red nodes has quite high degrees, and connects to many nodes in the network, while the remaining nodes has quite small degrees. In 1999, American young physicist Albert-Laszlo Barabasi and his student Reka Albert^[7] found that the degree distribution of such heterogeneous networks follows power-law distribution: $P(k) \propto k^{-\gamma}$ with power-law exponent as γ . They call this form of distribution network as scale-free

network and the power-law distribution degree of degree of network node as scale-free characteristic.

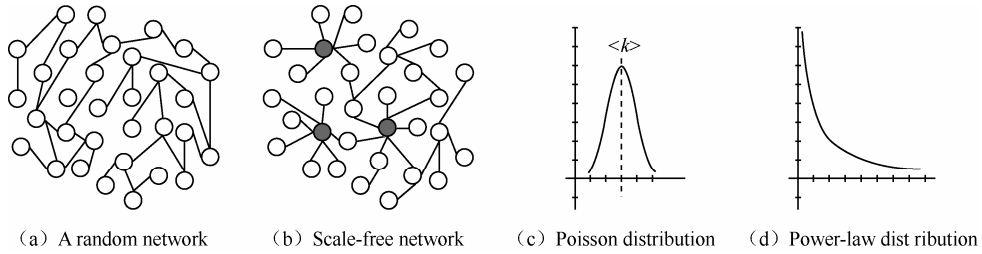


Figure 2-3 Random networks and scale-free networks and their distribution characteristics

Heavy-tailed phenomenon is an important characteristic of power-law distribution in scale-free networks. As shown in Figure 2-3(d), power-law distribution doesn't have node value as Poisson distribution and drags a long tail curve. Degree distribution in online social networks generally have heavy-tailed phenomenon. Unlike Poisson distribution, probability $P(k)$ that a node with degree of k appears in power-law distribution does not decrease at an exponential rate while k increases, but more gentle and progressively tend to 0, which shows the "long tail" nature. Long-tailed distribution is a subclass of heavy-tailed distribution, first pointed out by the Pareto Principle and Zipf's Law. In 1897, Italian economist Vilfredo Pareto studied the wealth and income patterns of British in the 19th century, and found that the incomes of few persons are far more than that of other persons, thus promote the famous "80/20 Rule", that is, 20% of the population accounted for 80% of social wealth. In 1932, George Kingsley Zipf, a linguistics expert from Harvard University, studied the usage frequency of English words, and found that the usage frequencies of words are not uniform when arranging the word in descending order, but follow simple inverse relation with the power function by its ranking, suggesting that only a handful of English words are often used while the vast majority of words are rarely used. In many real networks, the nodes with high degrees are rare compared with total number of nodes, but they played a "leading" role. And these degree nodes give the network the nature completely different from uniform random network.

Random network model^[8] assumes that the probability that any pair of nodes is interconnected is equal, and its degree distribution $P(k)$ follows Poisson distribution. When node degree k tends to infinity, the speed that Poisson distribution $P(k)$ tends to 0 is between normal distribution e^{-k^2} and exponential distribution e^{-k} . While the speed that exponential distribution e^{-k} tends to speed 0 is already fast, the speed that Poisson

distribution tends to speed 0 is much faster. But overall, these three distributions are “narrow tail” or almost “no tail”. We define the heavy-tailed phenomenon as follows, if the random variable X satisfies:

$$\lim_{x \rightarrow \infty} \frac{P(X > x)}{e^{-\lambda x}} = \infty \quad (2-9)$$

Then X have heavy tails. Wherein, λ is a positive integer. Heavy-tailed phenomenon means that, when tends to positive infinity, the probability for X taking $X > x$ is low order infinitesimal following exponential distribution, which indicates a higher probability than that of exponential distribution when the variable value is large. Obviously, for exponential distribution $p(x) = \lambda e^{-\lambda x}, x > 0$, substituted it into Formula (2-9) for verification:

$$\lim_{x \rightarrow \infty} \frac{P(X > x)}{e^{-\lambda x}} = \lim_{x \rightarrow \infty} \frac{e^{-\lambda x}}{e^{-\lambda x}} = 1 \neq \infty$$

Therefore, the exponential distribution does not have heavy tails^[9].

According to the knowledge of mathematical analysis, it is easy to know that the variable X following power-law distribution $P(x) = cx^{-\gamma}$ has heavy tails. In addition, a class of distribution frequently appears in real networks, and has heavy tails, power-law distribution truncated exponential, i.e. the variable X follow the distribution of $P(x) \propto x^{-\gamma} e^{-\lambda x}$.

After Albert-Laszlo Barabasi and Reka Albert^[7] published a landmark article in “Nature” and proposed the scale-free network model, they analyzed the degree distribution law of many real complex networks, such as the movie actor/actress collaboration network, the World Wide Web, the power grid of western United States, etc. They found that all these works similarly or exactly follow power-law distribution $P(k) \propto k^{-\gamma}$, and the power-law distribution exponent met $2 < \gamma < 3$, in which γ is a positive number. Their work reveals that important nodes (nodes with big degrees) in the actual networks distribute non-uniformly but orderly. Figure 2-4(a) indicates the degree distribution of the actor/actress cooperation network, in which the nodes denotes the actors/actresses and the edges denotes the cooperation relationship between the actors/actresses. The number of nodes in the network is 212,250 with an average degree of 28.78. Take logarithm for two sides of equation like $P(k) = ak^{-\gamma}$ and obtain $\log P(k) = \log a - \gamma \log k$, that is, if a power-law relationship exists, the function taking $\log k$ as the variable will be a straight line with a slope of $-\gamma$ in double logarithmic coordinate. Degree distribution network of actor/actress cooperation presents a power-law distribution $P(k) \propto k^{-\gamma_{\text{actor}}}$ with the exponent as $\gamma_{\text{actor}} = 2.3 \pm 0.1$. In World Wide Web, nodes are web pages and edges are

hyperlinks relationship between web pages. As shown in Figure 2-4 (b), the dataset is a subnet of World Wide Web containing 325,729 nodes and follows power-law distribution $P(k) \propto k^{-\gamma_{\text{www}}}$ with the exponent as $\gamma_{\text{www}} = 2.1 \pm 0.1$. In electricity network, nodes are motors, transformers or substations and edges are transmission lines between them. As shown in Figure 2-4(c), the power grid of western United States has 4,941 nodes and its degree distribution follows power-law distribution $\gamma_{\text{power}} = 4$.

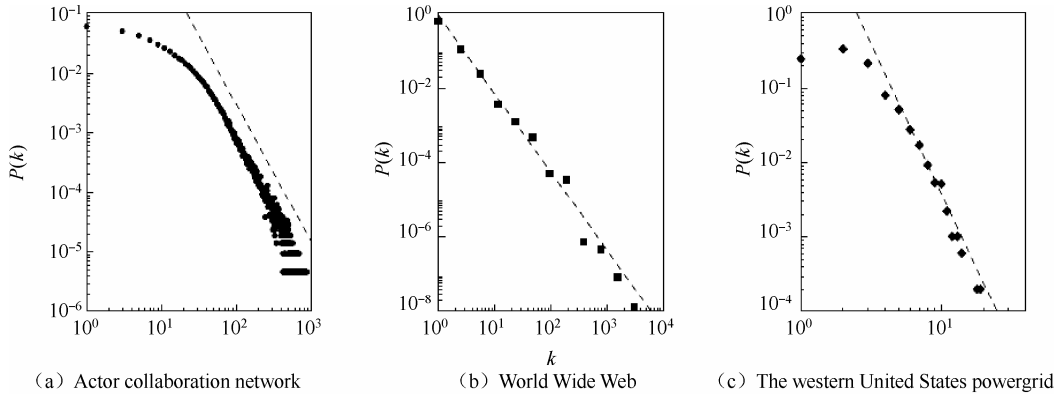


Figure 2-4 Diagram of network degree distribution^[7]

Many actual measurement studies on online social network show that social networks have scale-free characteristic. Most users have relatively few social relations while some users have more social relations, and the degree distribution follows power-law distribution. Alan Mislove et al.^[3] studied topological characteristics of largest connected subgraphs of 4 online social networks (Flickr, LiveJournal, YouTube and Orkut). Flickr is a website mainly for photo-sharing with an important characteristic of the extension of personal relationships and content organization based on social network. LiveJournal is an online community providing comprehensive social services including diary, blog, forum, dating. YouTube is a website mainly for video sharing where users can share videos on the website and download movies or short film and users can form a network community through friendships. Orkut, launched by Google, is an online community mainly for making friends, and Orkut users can leave their personal or professional information to create a relationship between friends or join the virtual community due to the same interest.

There are statistical results of 4 kinds of online social network in Table 2-7. Registered users are defined as nodes, and each user's friend lists are defined as the edge from the node to other nodes, making the entire network graph a directed graph. These 4 kinds of

online social networks impose no restrictions on access customer relationship. Flickr, LiveJournal and YouTube provide developers with open APIs to retrieve structured data, while social relationships in Orkut may be obtained only through web capturing. In addition, Orkut also limits the data scale that may be captured when accessing user account and login IP, making its data capturing scale smaller than the remaining three kinds of online social networks. Figure 2-5 shows the CCDF graph of the 4 kinds of online social networks. It is easy to see that the degree distribution of 4 kinds of distributed online social networks follow power-law distribution. In out-degree distribution of Orkut and LiveJournal, inflection point appears due to their limit on the number of friends in each node. In addition to limit on the number of friends, data capturing method of BFS algorithm make the samples from nodes with low degree too few, resulting in possible measurement bias in Orkut degree distribution.

Table 2-7 Statistical results of 4 kinds of online social networks

	Flickr	LiveJournal	Orkut	YouTube
Nodes	1,846,198	5,284,457	3,072,441	11,578,872
Capturing percentage	26.9%	95.4%	11.3%	Unknown
Number of edges	22,613,981	77,402,652	223,534,301	4,945,382
Average node degree	12.24	16.97	106.1	4.29

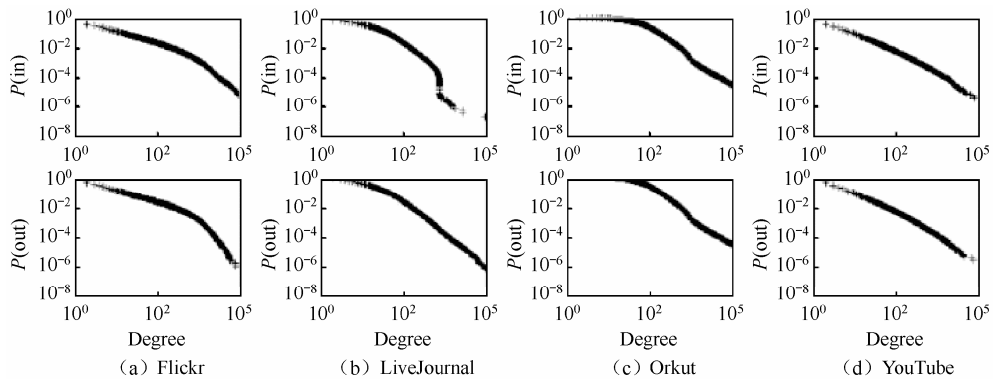


Figure 2-5 CCDF graph of accumulative degree distribution of 4 kinds of online social network^[3]

We use maximum likelihood method^[10] to fit curves in Figure 2-5 and the results are as shown in Table 2-8, where Δ denotes fitting bias. Overall, there is big difference between the exponents of out-degree and in-degree power-law distribution of Web, while the exponents of out-degree and in-degree power-law distribution of online social networks are

similar and less than that of Web as most friend relations in online social networks are bidirectional. Degree distribution of some online social networks is not monotonous power-law distribution. For example, Haewoon Kwak^[11] et al. found that in Twitter, an online social network like microblog, in class online, some users have a large number of “fans”, and “fans” distribution does not follow power-law distribution. Besides, Twitter, as an online social network for resource and information sharing, has a number of star nodes that has a lot of fans.

Table 2-8 Power-law exponent and bias of the degree distribution of several online social networks and Web^[3]

Online social network	Out-degree		In-degree	
	T	Δ	T	Δ
Web	2.67	—	2.09	—
Flickr	1.74	0.057 5	1.78	0.027 8
LiveJournal	1.59	0.078 3	1.65	0.103 7
Orkut	1.50	0.631 9	1.50	0.6203
YouTube	1.63	0.131 4	1.99	—

In scale-free networks, nodes with larger degree usually have smaller clustering coefficients, while nodes within smaller degree have larger clustering coefficients. In many real networks, the relation of clustering coefficient and degree follows power-law: $C(k) \propto k^{-\alpha}$, where $C(k)$ is the average clustering coefficient of nodes with degree of k and α is “level exponent”, then the network hierarchy exists^[12], i.e. the network can be clearly classified into several distinctive levels. Erzsebet Ravasz and Albert-Laszlo Barabasi found that scale-free networks have different hierarchies with close connection inside the group, but the average degree of nodes is small with sparse connection inside the group while hub nodes responsible for connections have relatively big degree. They also pointed out that the average clustering coefficients of networks like actor/actress cooperation network and the Internet substantially accords with decreasing power-law relationship.

Previous studies have shown that the clustering coefficient distribution of many online social networks follow power-law distribution^[13,14] with hierarchical structure. As shown in Figure 2-6, Cyworld^[4], an online social network in South Korean, has an average clustering coefficient of 0.16, slightly less than other online social networks^[3,13], indicating that relations between friends in Cyworld are relatively sparse. In addition, the clustering

coefficient distribution also has obvious segment characteristics. The clustering coefficient meets the power-law distribution with exponent of 0.4 when $k < 500$ and sharply decline when $k > 500$, i.e. connections between neighbors of nodes with degree value more than 500 are much sparser than that of those nodes with low degree. In general, nodes with small degrees in social networks have relatively high clustering coefficients and follow power-law distribution with some degree of hierarchy and relatively high aggregation, while nodes within the larger lower degree have low clustering coefficients and sparser distribution.

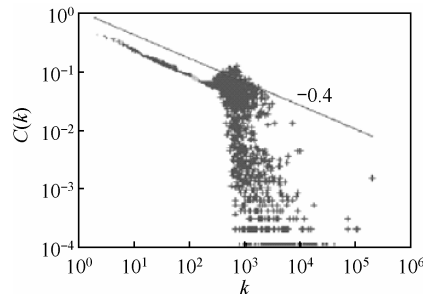


Figure 2-6 Clustering coefficient distribution of Cyworld^[4]

2.4.3 Assortativity

Assortativity reflects the correlation degree among nodes with similar degrees in the network. Among them, the degree correlation indicates the correlation between the degree of a node and its neighbor nodes. In a network, if node with larger (smaller) degree tends to connect a node with larger (smaller) degree, the network has positive correlation, i.e. assortativity, otherwise the network has negative correlation, i.e. disassortativity. We generally use two methods to measure the assortativity of the network: one is to draw a neighbor's average degree distribution and calculate the slope, the other is to calculate assortativity coefficient of the network.

1. Distribution of Average Neighbor Degree

Usually, the distribution of average neighbor degree is calculated through degree correlation function k_{nn} , which is defined as average value of average neighbor degree of a node with degree of k , and calculation formula is

$$k_{nn}(k) = \sum_k (k'P(k'|k)) \quad (2-10)$$

where conditional probability $P(k'|k)$ is the probability that there are edges between nodes with degree of k and k' . Actually, when calculating k_{nn} , we usually choose average neighbor degree of one node to replace it. First, average neighbor degree of node v_i is defined as

$$k_{nn,i} = \frac{1}{k_i} \sum_j a_{ij} k_j \quad (2-11)$$

where k_i denotes the degree of node v_i , a_{ij} is an adjacent matrix element. If v_i and v_j are connected, $a_{ij}=1$, or $a_{ij}=0$. So the average value $k_{nn}(k)$ of average neighbor degree of all nodes with degree of k is defined as

$$k_{nn}(k) = \frac{1}{|M_k|} \sum_{i \in M_k} k_{nn,i} \quad (2-12)$$

where M_k denotes the set of nodes with degree of k , $|M_k|$ is the node number of the set. In scale-free network, nodes with larger degree are minority, so a node with larger degree has smaller average neighbor degree. If $k_{nn}(k)$ is an increasing function with increasing of k , then it indicates that nodes with larger degree tends to connect with each other and such network is assortative network, otherwise it is disassortative network.

Example 2-4 Calculate the distribution of average neighbor degree of the following network in Figure 2-1 (b), and judge the network type.

Solution: Substitute degree sequence of each node listed in Table 2-2 into Formula (2-11) to calculate average neighbor degree of nodes:

$k_{nn, G} = \deg(E) = 5$, $k_{nn, C} = [\deg(E) + \deg(F)]/2 = 11/2$, In a similar way, we can calculate average neighbor degree of other nodes:

$$k_{nn, E} = 14/5, k_{nn, B} = 14/3, k_{nn, A} = 14/3, k_{nn, F} = 17/6, k_{nn, H} = 4, k_{nn, D} = 4.$$

Then substitute average neighbor degree of all above nodes into Formula (2-12), we can obtain:

$$k_{nn}(1) = k_{nn, G} = 5$$

$$k_{nn}(2) = (k_{nn, C} + k_{nn, H} + k_{nn, D})/3 = 9/2$$

In a similar way, we can obtain other $k_{nn}(k)$, then the distribution of average neighbor degree is as shown in the following Table 2-9.

Table 2-9 Distribution of average neighbor degree of the network shown in Figure 2-1 (b)

k	1	2	3	4	5	6
$k_{nn}(k)$	5	9/2	14/3	0	14/5	17/6

In Table 2-9, $k_{nn}(k)$ tends to decline with increase of k , so the example network in Figure 2-1(b) is in negative correlation, thus the network is disassortative network.

Most social networks are assortative networks, such as scientist cooperation network and movie actor/actress cooperation network^[19]. Most scientists and movie actors/actresses hope to cooperate with persons at the same level, so their social network is assortative network. However, many online social networks are disassortative networks. Online social networks shorten the distance among persons, which makes every ordinary people has a chance to easily build one-way friend relationship with “star node” with larger degree. Figure 2-7 describes the distribution situation of average neighbor degree in Cyworld, an online social network in South Korea, which represents complex non-monotonous characteristic, while we can see that overall trend of the average neighbor degree distribution curve is in negative correlation. Yong-Yeol Ahn believes that there are different types of users in social networks, and the combination of different types of users leads to the complex distribution of average neighbor degree.

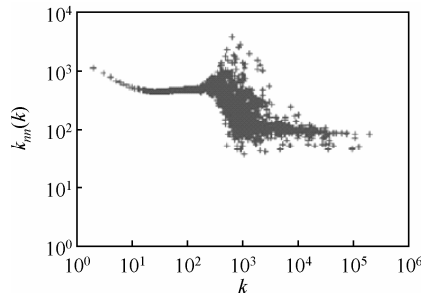


Figure 2-7 Distribution situation of average neighbor degree of Cyworld, an online social network^[4]

Alan Mislove et al.^[3] found that the distribution function of average neighbor degree of Flickr, LiveJournal and Orkut tend to increase while that of YouTube tends to decrease. This variation trend indicates that, in YouTube, nodes with larger degrees are more likely to connect those with smaller degrees. As a social network site for video sharing, YouTube certainly has some very popular resource sharing users who can attract other users to connect, and become the star nodes of the network, which is consistent with the characteristics of video sharing. We can roughly determine the assortativity of the network through the trend of the distribution function of average neighbor degree; however, some distribution trends of average neighbor degree in the online social network may be more complex than that is shown in Figure 2-7. We will introduce a calculation method below to determine the assortativity of network by substituting the structure information of each node into the

formula.

2. Assortativity Coefficient

The assortativity coefficient of node r is used to evaluate the relationship between the degree of a node in a network and the degree of its adjacent nodes^[19]. The definition is as follow:

$$r = \frac{M^{-1} \sum_i j_i k_i - \left[M^{-1} \sum_i \frac{1}{2} (j_i + k_i) \right]^2}{M^{-1} \sum_i \frac{1}{2} (j_i^2 + k_i^2) - \left[M^{-1} \sum_i \frac{1}{2} (j_i + k_i) \right]^2} \quad (2-13)$$

where k_i and j_i denote the degree of two nodes connecting with the i edge. M denotes the total number of edges in the network. If the value of r exceeds 0, the degree of the mutual neighbor nodes in the network is in positive correlation. Nodes with larger degree tend to connect with each other while the adjacent nodes of nodes with smaller degree generally have smaller degrees, thus such networks have assortativity the corresponding correlation coefficient is called assortativity coefficient. The range of assortativity coefficient is easy to be proved as $0 \leq |r| \leq 1$. When $r < 0$, the network is disassortative; when $r > 0$, the network is assortative; when $r = 0$, the network is irrelevant, such as the random network.

Substitute the degree sequence of each node in Table 2-2 into Formula (2-13). Its calculation process is as follows:

$$r = \left[\frac{151}{12} - \left(\frac{92}{24} \right)^2 \right] / \left[\frac{420}{24} - \left(\frac{92}{24} \right)^2 \right] = -0.75$$

In Figure 2-1(b), the assortativity coefficient of Sina Weibo following network is less than zero ($r < 0$), indicating that the internet is disassortative. This is consistent with our previous conclusion drawn by using the distribution of average neighbor degree.

In the early study of complex networks, Mark Newman et al. took the lead in finding that there is assortativity in human social networks, which is different from other biological networks or science and technology networks with disassortativity. Common sense can easily explain the phenomenon: In real life, in the needs of social resources, people often want to expand their social circles. Limited to the individual's social classes, it is easy for elites to know each other while ordinary persons can only know similar persons. However, due to the convenience and low cost in social activities in online social networks, which breaks the constraints of human social classes, it is more easy for ordinary people to establish a one-way following or acquaintance with the elites. As related research confirms

the law, people find that online social networks generally have characteristics of assortativity or inconspicuous disassortativity (as shown in Table 2-10).

Table 2-10 Assortativity coefficients of online social networks

Network	Number of nodes	Assortativity coefficient	Reference	Network	Number of nodes	Assortativity coefficient	Reference
Cyworld	12,048,186	-0.13	[4]	Flickr	1,846,198	0.202	[3]
Nioki	50,259	-0.13	[22]	LiveJournal	5,284,457	0.179	[3]
MySpace	100,000	0.02	[4]	YouTube	1,157,827	-0.033	[3]
Orkut	100,000	0.31	[4]	Mixi	360,802	0.1215	[3]
Xiaonei	396,836	-0.0036	[16]				

Online social networks' assortativity has an evolutionary process. In the early stage after establishment, social networks usually have assortativity; however, along with the continuous increase of user group scale, many networks evolve into disassortativity from assortativity. This is because the early users of social networks are usually introduced by other people who first joined it, such user can restore offline "face-to-face" social relationships vividly. In this stage, the degree correlation of a network node degree represents higher assortativity. However, with the continuous extension and development of the network scale, the site can attract well-known figures, and the original network begins to have excellent users like "opinion leaders", which makes a lot of ordinary users with lower node degree choose to connect these elite users, thus the network evolves into a disassortativity network^[17]. Neil Gong et al.^[18] studied the link relationship data of the online social network Google+ in several months from the close beta test to open, and found that the assortativity coefficient follows the revolution rule of changing from positive to negative. They think Google+ is a hybrid network consists of two different types of networks, i.e. traditional dating social networks and publish-subscribe networks like YouTube, Sina Weibo, etc. The former usually has a positive assortativity coefficient while that of the latter may be negative. In the beginning of close beta test, the traditional dating social networks have superiority but the two different types of networks gradually integrated with each other along with lapse of time and increase of users, finally the publish-subscribe networks dominates.

2.4.4 Reciprocity

Reciprocity is generally used in a directed network to measure the extent that two nodes in the network form a two-way connection^[20]. Research on network reciprocity is of

good guidance. On one hand, reciprocity may reflect the closeness of the interaction between individuals in the network; on the other hand, for the need of simplicity in the actual operation, we often ignore the direction of directed edges while reciprocity can reveal errors caused by ignoring the direction of the edges.

We can use mutual reciprocity coefficient to quantify reciprocity with its mathematical representation as $\varphi = m_d / m$ where m denotes the total number of edges in the network and m_d denotes the number of edges with reverse edges. Real meaning of reciprocal coefficient is very intuitive, i.e. randomly choosing a directed edge from node A to node B from the directed network, then possibility that there is a directed edge from node B to node A. In Figure 2-1(a), Kun Chen \rightarrow Chen Yao, Zhiying Lin \rightarrow Degang Guo are unidirectional following relations while the remaining are two-way following relations, so the reciprocity coefficient of the network is $(m - 2) / m = 5 / 6$.

The reciprocity coefficient in dating online social networks are usually higher, for example, reciprocity coefficient of LiveJournal and Flickr is respectively 0.74^[4] and 0.68^[21]. In sharing-based microblog networks, there are many celebrity and media nodes with a lot of fans (in-degree) but rare following (out-degree), so reciprocity of microblog are usually poor, for example, reciprocal coefficient of Twitter is only 0.22^[11].

2.5 Social Network Structure Modeling and Generation

For social network characteristics discussed in Section 2.4, people often use structure modeling to study the network evolution mechanism of these characteristics. For example, WS model (see Section 2.5.1) introduces randomness for regular network by reconnection mechanism, and indicates that such social networks as actor/actress cooperation network and scientific research reference network are substantially a kind of complex between regular network and random network, i.e. small-world network. Through simulating the phenomenon in WWW network that minority web pages indexed by a large number of other web pages, BA model (see Section 2.5.3) promotes the evolution mechanism that nodes tend to establish connections by “preferential attachment” law, and reveals the reason of scale-free characteristics in network. The research on network model is helpful in understanding the formation process of social network and reason for generating some special phenomena, and further deepens our understanding of internal law and substantial characteristic of social network.

As a typical complex network, social network, in its network structure and without

loss of generality, represents small-world phenomenon and scale-free characteristic of complex network as well as assortativity, reciprocity and other characteristics brought by human's social behavior. Through specifically description on network formation process and corresponding instances, this section will thoroughly discuss small-world network model and scale-free network model, and provide an overview of other important networks in Section 2.5.5.

2.5.1 WS Model

In 1998, Duncan Watts and Steven Strogatz presented the concept of small-world network and established small-world model, i.e. WS model^[24]. As described in Section 2.4.1, the small-world phenomenon reveals characteristics of many complex networks in objective world, i.e. larger average clustering coefficient and shorter average path length. Among them, the shorter average path length is realized through long-distance connection (long-distance connection refers to the connection formed between two nodes with relatively long distance) formed during edge re-connection process.

1. Description of Algorithm

Small-world network formed through WS model is the intermediary status of network during the transition from regular network to random network. A network is generated from WS model in the following steps^[25]:

(1) We use an annular grid network including n node(s) with node degree of $2k$ as the initial network, and each node in the network is connected to $2k$ node(s) most adjacent to it. Among them, k is an integer larger than zero (usually with small value).

(2) We specify a probability p and reconnect to each edge in the initial network at the probability of p (when reconnecting, randomly choose a node to replace a node connected to such edge). New connection shall not be self-connection and repetitive-connection.

In Step (2), edge connection will generate long-distance connection between two nodes. According to the above process, when $p=0$, the graph obtained is still original regular network; when $p=1$, each edge in original graph is randomly re-connected to finally form an approximate random network; when $0 < p < 1$, the original regular network gradually evolves into a small-world network to form an approximate random network.

2. Algorithm Example and Network Generation

Currently, most tools provide the algorithm implementation of WS model, such as large-scale complex network analysis tool Pajek^①, social network research tool SNAP^② of Stanford University, R language or igraph^③ under Python environment, etc. The following is code for generating small-world network using SNAP tool with given parameters, and corresponding network structure graph and some statistical characteristic graphs are generated through other integrated open-source tool, including drawing tool Gnuplot and visualization tool Graphviz. After proper environmental configuration, generation and analysis of the network can be completed by inputting the following code into Python interactive interpreter.

```
1 import snap
2 Rnd = snap.TRnd(1,0)
3 UGraph = snap.GenSmallWorld(N,k,p,Rnd)
4 snap.DrawGViz(UGraph, snap.gvlCirco, "WS.png", "WS small world", False)
5 GraphClustCoeff = snap.GetClustCf(UGraph,-1)
```

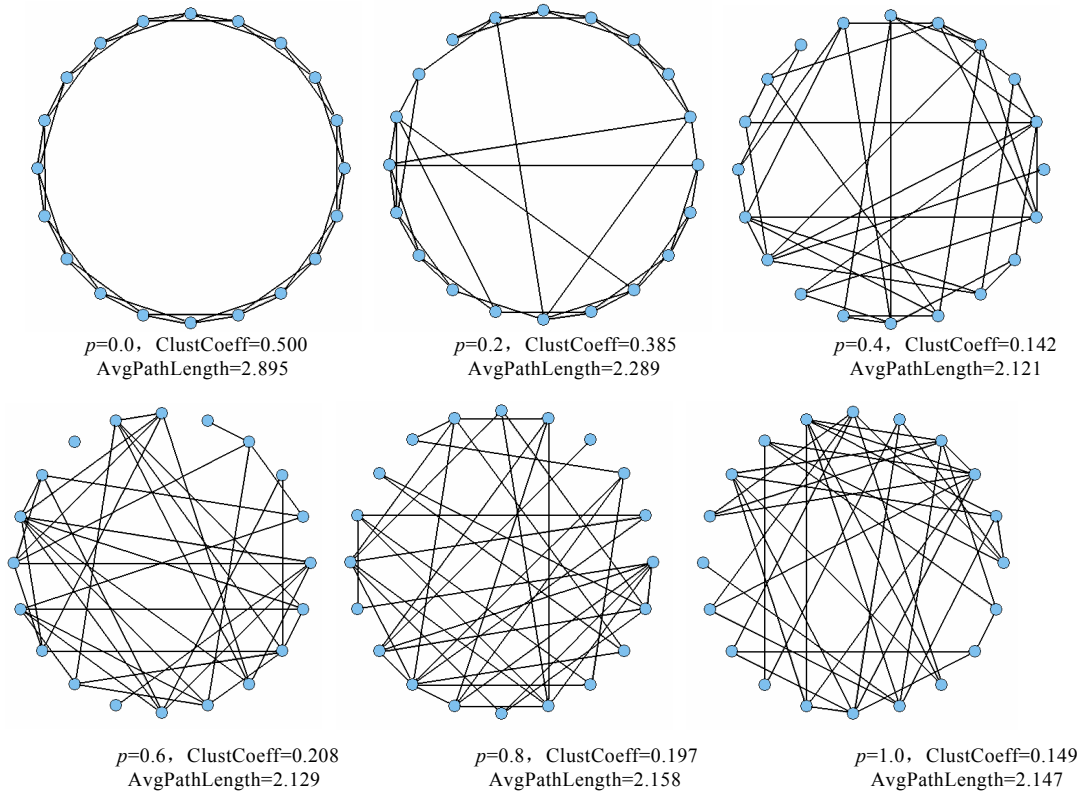
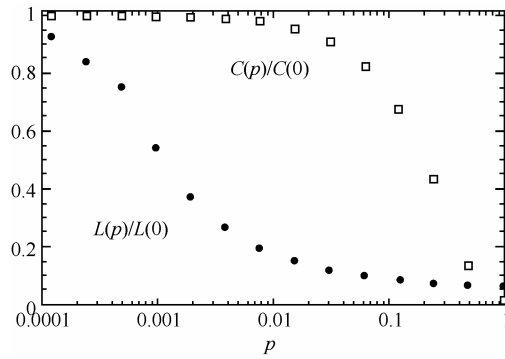
where the first line of code is for importing SNAP toolkit; the second line of code generates a random number generator; the third line of code generates a small-world network named as UGraph; the fourth line of code draws UGraph network under gvlCirco layout by invoking open-source tool DrawGViz and save it into a png file named as WS by current operating path, with parameter False indicating no labeling of node serial number in the network. By adjusting parameter N , k and p in function GenSmallWorld(), scale of the generated network, half of average degree and probability of edge re-connection can be limited accordingly. As shown in Figure 2-8, when $N=20$, $k=2$, a group of WS network examples are generated according to different values of p .

where ClustCoeff denotes average clustering coefficient, AvgPathLength denotes average path length. From the first graph, WS small-world network evolves from regular annular grid network to random network as probability of edge re-connection increases from 0 to 1, and average path length changes accordingly as average clustering coefficient decreases. However, we cannot obtain the change law of average clustering coefficient and average path length through just 6 samplings. The evolution law of average clustering coefficient and average path length after multiple samplings is as shown in Figure 2-9.

① <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>.

② <http://snap.stanford.edu/data/>.

③ <http://igraph.org/redirect.html>.


 Figure 2-8 WS network generated according to different values of p

 Figure 2-9 Evolution law of average clustering coefficient and average path length^[24]

where $L(0)$ and $C(0)$ respectively denote average path length and average clustering coefficient of regular network, $L(p)$ and $C(p)$ respectively denote average path length and average clustering coefficient of the network obtained by edge re-connection at probability p . $L(p)/L(0)$ and $C(p)/C(0)$ respectively denote the normalization process to $L(p)$ and $C(p)$

by $L(0)$ and $C(0)$. As shown in Figure 2-9, the two parameters show rather different decreasing ratio and law as p increase from 0 to 1: the average path length decreases sharply while average clustering coefficient decreases relatively slow. Therefore, with proper value of p , the network can have relatively small average length while keeping relatively high average clustering coefficient.

2.5.2 Extension of WS Model

Though small-world network with relatively high average clustering coefficient and average path length can be generated in WS model, edge re-connection step therein may harm the connectivity of network and generate some disconnected branches. To remedy this defect, Mark Newman and Duncan Watts modify the “random edge re-connection” to “random edge addition”^[26] and maintain the connectivity of network by slight modification on original model. The modified model is named as NW model.

1. Description of Algorithm

The establishment process of NW model is as follows:

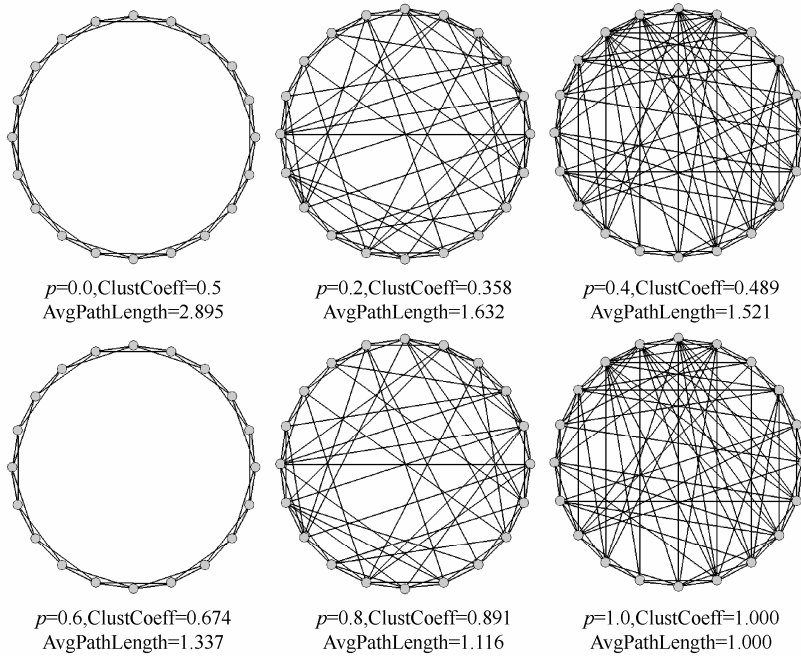
(1) Similarly, we use an annular grid network including n node(s) as the initial network, and each node in the network is connected to $2k$ node(s) most adjacent to it. Among them, k is an integer larger than zero (usually with small value).

(2) For each pair of disconnected node (i, j) in the initial network, add an edge e_{ij} between node i and j at probability of p .

In this process, there is one edge between each node pair at most without self-loop. In this model, when $p=0$, the generated network is still annular grid network; when $p=1$, a complete graph forms; when $0 < p < 1$, the sparse regular network evolves into small-world network and finally forms a dense regular network. When network scale n is large enough and p is small enough, the network generated from NW model is basically same to that of WS model.

2. Algorithm Example and Network Generation

NW model may be realized through the method similar to WS model. As shown in Figure 2-10, when $n=20$, $k=2$, a group of NW network examples are generated according to different values of p .

Figure 2-10 NW network generated according to different values of p

Similarly, ClustCoeff denotes average clustering coefficient, AvgPathLength denotes average path length. From the first graph, as probability of edge addition increases from 0 to 1, the network formed has bigger average clustering coefficient and smaller average path length, and the NW small-world network becomes denser and denser. Finally, the initial sparse regular network (annular grid network) forms a dense regular network through small-world network (complete graph).

Besides WS model and NW model, small-world network has other model like Monasson small-world network model^[27], BW small-world network model^[28] and so on. Please refer to relevant reference.

2.5.3 BA Model

In traditional random network and small-world network, node degree distribution follows bell-shaped Poisson distribution, i.e. reaching peak around average and decrease exponentially at two sides. However, according to research on networks in real world, node degree therein follows power-law distribution, i.e. for a randomly specified node in the

network, the probability $p(k)$ that it has degree of k follows $p(k) \sim k^{-\gamma}$, which indicates that large-scale network will self-organize into a scale-free status. Scale-free here refers to lack of characteristic degree (average degree) in the network, which leads to big fluctuation range. Albert-Laszlo Barabasi and Reka Albert give full explanation upon such phenomenon by simulating resource concentration effect in network through “preferential attachment” mechanism in BA model. We will detail BA model, an important research result of network structure model research, and other improved extension models.

1. Description and Analysis of Algorithm

BA model takes into account that network generally has the phenomenon of scale expansion and “the rich becomes richer”, i.e. nodes with higher degrees tends to be connected by other nodes and thereby occupy more network resources (higher node degree), and promotes two major factors for the network self-organization in scale-free structure.

(1) Network expansion: Network scale is always expanding. For a fully connected network with scale of m_0 at the beginning, add a new node each time and connect it to $m(m \leq m_0)$ existing nodes.

(2) Preferential attachment: The probability Π_i that new nodes connect existing node i depends on the degree k_i of node i , i.e. $\Pi_i = k_i / \sum_j k_j$.

(3) Repeat Step (1) and (2) until the network scale reaches N .

After time t , there will be $N = t + m_0$ node(s) and mt edge(s) in the network. When t is large enough, m_0 can be omitted to infer that the degree distribution of BA model follows $p(k) \approx 2m^2 k^{-3}$, i.e. power-law distribution. Through mathematical inference, we can obtain the average path length $L^{[29]}$ and average clustering coefficient $C^{[30]}$ of scale-free model:

$$L = \frac{\ln(N)}{\ln \ln(N)}$$

$$C = \frac{m^2(m+1)^2}{4(m-1)} \left[\ln \left(\frac{m+1}{m} \right) - \frac{1}{m+1} \right] \frac{[\ln(t)]^2}{t}$$

In regular network, random network and scale-free network, average path length and average clustering coefficient has different change laws along with the increase of network

scale: the increase of average path length is fastest in regular network, middle in random network and slowest in scale-free network; for average clustering coefficient, regular network remains the same, random network decreases fastest and scale-free network is at the middle.

BA model generates scale-free network by network expansion and “preferential attachment”, and lays the foundation for other scale-free network models. Actually, most scale-free network models are the modification or extension version of BA model. The finding of scale-free characteristics and promotion of BA model also create an upsurge of network science research.

2. Algorithm Example

To better present the evolution mechanism of BA model, we detail BA model through the generation process of a small-scale network with 8 nodes as follows.

Figure 2-11 shows how BA model generates a network, with model parameter as $m_0 = 3, m = 2$. There are 8 steps as follows from upper-left to lower-left:

(1) At the beginning, establish a fully connected network with 3 nodes.

(2) Add a new node in the network.

(3) Connect the new node randomly to two nodes of the existing three nodes. The probability for connecting to each node is equal as existing nodes has the same degree, and can be calculated according to probability formula $\Pi_i = \frac{1}{3}$.

(4) Add another node in the network.

(5) Connect the new node to two existing nodes according to probability. For the two classes of nodes with degrees of 2 and 3 in Figure 2-11, assume that the probability to be connected are Π_2 and Π_3 respectively, then $\Pi_2 = \frac{2}{2+2+3+3} = \frac{1}{5}$,

$\Pi_3 = \frac{3}{2+2+3+3} = \frac{3}{10}$, i.e. the probability for connecting the new node to two nodes with

degree of 3 is higher. In the graph, the “preferential attachment” characteristic of BA model begins to show and become more obvious along with the increase of degree difference between nodes.

(6) Connect the new node to two nodes with the largest degree (the event with the highest probability).

(7) Connect the new node to two nodes with the largest degree (the event with the

highest probability).

(8) The probability for connecting the new node to a node with the largest degree is the highest, and to a node with the smallest degree is the lowest. In the subsequent evolution process, degrees of high-degree nodes tend to increase faster and low-degree nodes slower, with a few low-degree nodes evolving into high-degree nodes. The network generated according to the above evolution mechanism can turn into a scale-free network featured by power-law distribution and heavy-tailed phenomenon at the end of evolution.

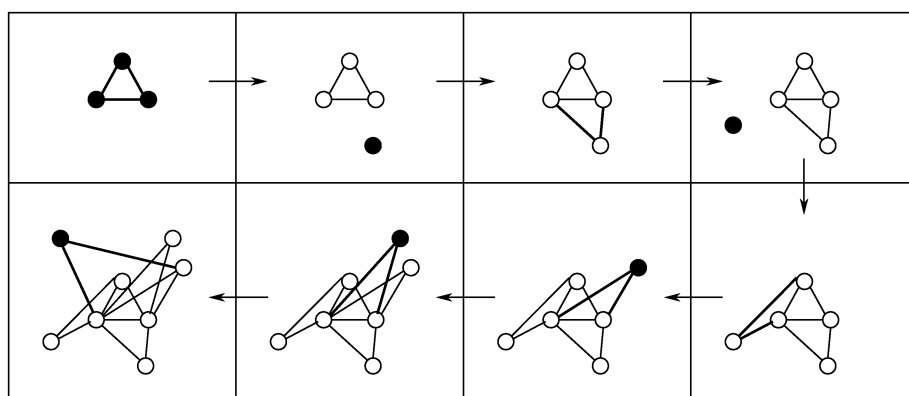


Figure 2-11 The network generation process by BA model

3. Generation of Instance

Use SNAP tool to generate BA network with certain scale, and generate network structure graph and statistical characteristic graph by other open-source tools, including drawing tool Gnuplot and visualization tool Graphviz. We will show a network with scale of $N = 500$ and average degree of $k = 1$. After the installation and configuration of tools, input the following code into Python interactive interpreter to realize network generation and analysis.

```
1 import snap
2 Rnd = snap.TRnd()
3 UGraph = snap.GenPrefAttach(500,1,Rnd)
4 snap.DrawGViz(UGraph, snap.gvlSfdp, "graph_generation_exp.png", "BA network", False)
5 snap.PlotInDegDistr(UGraph, "Degree distribution", "Degree distribution(BA network)")
```

#导入工具包
#生成随机概率
#产生BA网络对象
#网络绘图
#属性绘图

where the first line of code is for importing SNAP toolkit; the second line of code generates a random number generator; the third line of code generates a undirected BA network named as UGraph; the fourth line of code draws UGraph network under gvlCirco layout by invoking DrawGViz and saves it into a png file named as graph_generation_exp

by current operating path, with parameter False indicating it does not label the serial number of node in the network (see Figure 2-12 for detailed network graph), and the fifth line of code draws the degree distribution graph by invoking open-source tool Gnuplot as shown in Figure 2-13.

As shown in Figure 2-12, several high-degree nodes occupy most degrees in the network generated by BA model. As shown in 2-13, degree distribution graph will be a straight line in double logarithmic coordinate, i.e. the generated network follows power-law distribution.

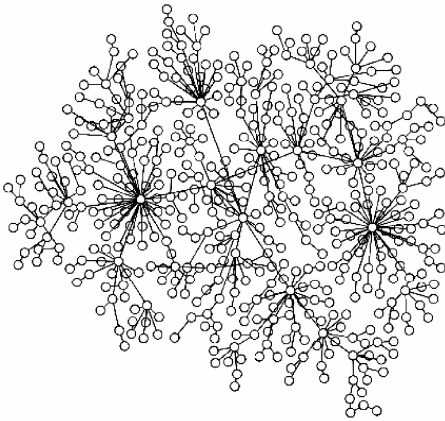


Figure 2-12 Network structure graph of BA scale-free network with 500 nodes

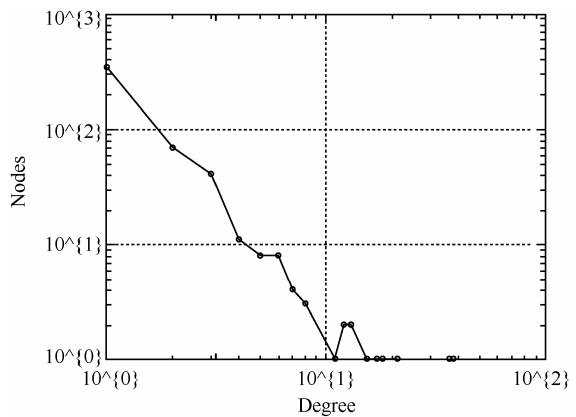


Figure 2-13 Degree distribution graph of BA network

2.5.4 Extension of BA Model

The major contribution of BA scale-free model is that it can precisely depict basic characteristics of most networks, i.e. the expandability of network scale and “preferential attachment” characteristic of new nodes. However, many networks have some special or rather complex characteristics, which have significant effects on actions in network and cannot be depicted in BA model. Therefore, based on BA model, several kinds of extended scale-free network are promoted according to different real networks. We will provide an overview of several important extension version of BA model, mainly about their principles and evolution processes. Please see corresponding reference for more details.

1. EBA Model

As described in Section 2.5.3, in BA model, degree distribution $p(k) \sim k^{-\gamma}$ has power exponent of $\gamma = 3$. However, according to data from real networks, many networks have power exponent of $2 < \gamma < 3$. Based on this situation, Reka Albert and Albert-Laszlo Barabasi extended the original BA model and incorporate “edge re-connection” into the extension version to put forward the EBA model^[31]. In this model, there are m_0 isolate node(s) at the beginning and one of the following three operations will be carried out each time.

(1) Add $m(m \leq m_0)$ edge(s) at probability p . For each edge, choose randomly for one end and at probability of $\Pi(k_i) = \frac{k_i + 1}{\sum_j (k_j + 1)}$ for the other end (preferential connection). Repeat such choosing process for m times to establish m edge(s).

(2) Reconnect m edge(s) at probability q . Each re-connected edge will be generated according to the following rules: Randomly choose a node i and randomly delete an edge l_{ij} connected to it; establish new connection between node i and node j' chosen according to the probability $\Pi(k_{j'})$ in Step (1); repeat this process for m times to form m re-connected edge(s).

(3) Add a new node at probability $1 - p - q$ and establish m edge(s) from this node, with each edge connected to the existing node i at probability $\Pi(k_i)$.

As shown in the above processes, EBA model comprises of three main processes: adding node, adding edge and re-connecting edge. Adding node and adding edge stimulate the network expansion and “preferential attachment” while re-connecting edge stimulates the power exponent controlling power-law distribution. By introducing this simple edge re-connection mechanism, networks generated from EBA model are able to better fit real networks.

2. Adaptability Model

The evolution process of BA model implicitly decides that earlier node has higher node degree, which is quite different from actual situations. For example, influential young people often have more followers in social network, and new innovative website often attract more users in Internet. These phenomena result from different importance of nodes in the network. Ginestra Bianconi, Albert-Laszlo Barabasi and other researchers introduce

the importance of nodes into BA model, which is called as adaptability, and form adaptability model^[32]. The detailed processes of adaptability model are as follows.

(1) Network expansion: The network has m_0 node(s) at the beginning, then adds a new node each time and assigns adaptability $\eta_i (0 < \eta_i < 1)$ for it.

(2) Preferential attachment: Connect the new node to m existing node(s), with the probability Π_i of connecting to an existing node i positively correlated to its node degree k_i and adaptability η_i .

$$\Pi_i = \frac{\eta_i k_i}{\sum_{j=1}^n \eta_j k_j}$$

where $n = m_0 + t - 1$ are all nodes existing in the network at $t - 1$.

It's easy to see that after t steps, there will be $n = t + m_0$ node(s) and mt edge(s) in the network. Adaptability model is basically the same as BA model except that the probability of “preferential attachment” is not totally relies on node degree by introducing the concept of adaptability. Therefore, if the new node has higher adaptability, it will have higher degree than existing node with lower adaptability during the process of network evolution.

3. Generalized Linear Precedence Model

Real network usually have shorter average path length and bigger average clustering coefficient, which are difficult to depict in traditional BA model. To depict average path length and average clustering coefficient on the premise of maintaining scale-free characteristics, Tian Bu et al.^[33] promoted generalized linear precedence (GLP) model, which is generated in the following steps.

Carry out one of the following operations in a connected network with m_0 node(s) and $m_0 - 1$ edge(s).

(1) Add $m(0 \leq p \leq m_0)$ new edge(s) in the existing network at probability $p(0 \leq p \leq 1)$ and connect one end of each edge to node i at probability

$$\Pi(k_i) = \frac{k_i - \beta}{\sum_j (k_j - \beta)}, \text{ with } k_i \text{ as the degree of node } i.$$

(2) Add a new node at probability $1 - p(0 \leq p \leq 1)$ and $m(0 \leq p \leq m_0)$ new edge(s) in the network, and choose the other end i at probability $\Pi(k_i) = \frac{k_i - \beta}{\sum_j (k_j - \beta)}$ for each edge.

where $-\infty < \beta < 1$ is an adjustable parameter, which denotes its tendency of preferentially connecting to a more welcome existing node. When $\beta < 1$, nodes with degree of 1 have a chance to obtain new connections. GLP model is able to better match power-law coefficient and average clustering coefficient in real networks.

2.5.5 Other Models

Promotion of BA model and WS model changed the dominate status of ER random graph in network modeling, and, along with the deep-going development of network research, people began to realize that complex characteristics in network cannot be depicted and described by a single model any longer. According to characteristics of different networks, people promote corresponding models to probe its formation mechanism. These models can be roughly classified into two classes according to the motives of modeling: one class is used to generate network with certain characteristics, such as the abovementioned WS small-world model, NW small-world model, forest-fire model and Kronecker graph model; the other class is structured for researching impacts of certain actions on network structure characteristics, such as BA model and its improved models as well as production model. We will provide an overview of these models below. Please refer to corresponding reference for details.

1. Forest-fire Model

Jure Leskovec and other researchers^[15] found that networks follow the law below during their research on reference network and coauthor network: in-degree and out-degree follow heavy-tailed distribution, relations between number of nodes and edges (density) follow power-law distribution and effective diameters of networks decrease along with lapse of time. To simulate and generate networks following these laws, they promoted forest-fire model, through which modeling for directed networks can be carried out. During the network evolution process, add one node in network each time and establish several directed edges from such node to existing nodes. When connecting edges, new node v randomly connects to a existing node w , which is called as representative node, then establish a directed edge at certain probability from node v to neighbors of w (including out-neighbor and in-neighbor), expanding layer by layer like forest fire.

Specifically speaking, define two parameters: forward flaming probability p and backward flaming probability r . G_t is an existing network at t , with G_1 having only

one node. When adding node v in the network at $t \geq 1$, establish the directed edge from node v to a node in G_t according to the following steps.

(1) First randomly choose a representative node w for node v and establish the directed edge from node v to node w .

(2) Generate a random number x according to binomial distribution with average of $(1-p)^{-1}$, and choose x node(s) for v from neighbors of w , where the proportion for choosing in-neighbor is r times smaller than that of out-degree. Assume the chose nodes as w_1, w_2, \dots, w_x .

(3) Establish the directed edge from node v to nodes w_1, w_2, \dots, w_x , and invoke Step (2) recursively for nodes w_1, w_2, \dots, w_x . During the process, the visited nodes cannot be visited repeatedly so as to avoid from being trapped in loop.

In this model, regardless of the location of representative node in the network, high-degree nodes are more likely connected to new node, causing heavy-tailed phenomenon to in-degree node of network formed. Through recursive invoking during connecting process, a new node forms many out-edges and has big degree, resulting in heavy-tailed phenomenon of out-degree. New node in model may establish directed edges through recursion of neighbors of representative nodes, forming community structure in network. New node will form many connections around the community of representative node, granting networks generated from forest-fire model many characteristics observed in real network: heavy-tailed distribution of in-degree and out-degree, power-law distribution of density, decrease of effective diameter of network along with lapse of time, etc.

2. Kronecker Graph Model

In generation models simulated on the basis of network properties, some models use product of matrix to simulate the expansion and evolution of adjacent matrix^[34,35]. Jure Leskovec and other researchers found out that Kronecker product operation of matrix can be used to generate network^[36], and verified in experiments that network generated from Kronecker graph model can well simulate the degree distribution of static network, proper value distribution and power-law distribution of diameter and density of dynamic network, etc. Mathematical characteristics of Kronecker product give network generated from Kronecker graph model sound analyzability.

Kronecker product is a kind of matrix product operation. Given matrixes $A=[a_{i,j}]$ in size of $n \times m$ and B in size of $n' \times m'$, then the Kronecker product of matrixes A and

B denote a matrix C in size of $(n \cdot n') \times (m \cdot m')$ as follows

$$C = A \otimes B = \begin{pmatrix} a_{1,1}B & a_{1,2}B & \cdots & a_{1,m}B \\ a_{2,1}B & a_{2,2}B & \cdots & a_{2,m}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}B & a_{n,2}B & \cdots & a_{n,m}B \end{pmatrix}$$

As shown in the above formula, unlike other matrix multiplication, Kronecker product of matrix is extended operation of matrix. Through defining Kronecker product between two graphs as the Kronecker product of their adjacent matrix, these graphs can be expanded into graphs with self-similarity as shown in Figure 2-14.

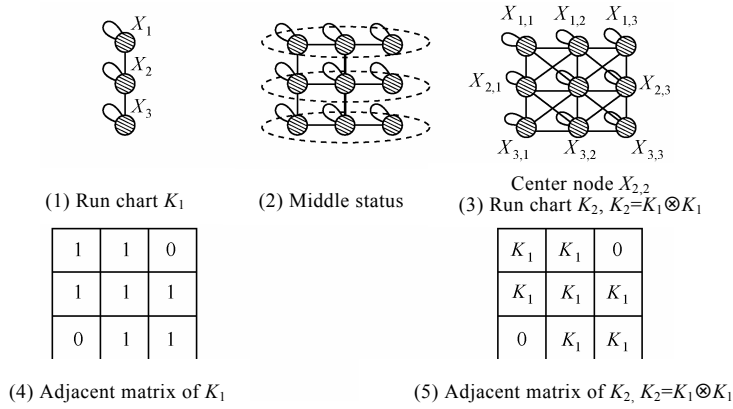


Figure 2-14 Kronecker product graphs

- (1) Run chart K_1 with three nodes.
- (2) Middle status of Kronecker product, indicating results after expansion.
- (3) Self-Kronecker product results of K_1 . K_i denote Kronecker products of i K_1 . Specially, K_2 denotes the Kronecker product of two K_1 .
- (4) Adjacent matrix of K_1 .
- (5) Adjacent matrix of K_2 .

Network generation process of Kronecker graph model is to carry out several times of Kronecker product operation and finally form K_i . It is easy to know that scale of K_i is i power of scale of K_1 . According to the mechanism of Kronecker product, even K_1 with a small scale, like 3×3 matrix, can finally generate a network with sound variability. To make the model better simulate networks in real world, the author promoted improved mode of Kronecker graph model, i.e. random Kronecker graph model, in which elements in K_1 adjacent matrix is replaced by probability value,

making Kronecker graph model more flexible. In this way, the improved model can generate networks with certain characteristics by changing parameters, and also simulate networks in real world by parameter evaluation.

3. Production Model

Production model was promoted by Ravi Kumar et al. during their observation and research on Flickr and Yahoo!360 network in 2006^[37]. It carries out modeling for directed network and classifies user nodes in network in two three classes: Passive, Linker and Inviter. Passive node may join the network out of curiosity or from continuous invitations from friends. But as implied by the name, Passive node acts passively and avoid from participating any activity in the network. Inviter node is devoted to moving offline communities to online activities and thereby continuously invites friends to join the network. Linker node is active participant in network activities and forwardly establishes relationship with other members.

During the network generation process, add one node and ε edge(s) in the network each time and randomly specify the node as one of Passive, Linker and Inviter. For edges, their classes are related to that of the node. Add each edge in the following steps: Choose a node as the edge source from existing Inviter node and Linker nodes in the network according to “preferential attachment” rule; in case of Inviter node, add a node in the network as the edge terminal for connection; in case of Linker node, choose a node as the edge terminal from existing Inviter and Linker nodes according to “preferential attachment” rule.

Network generated by this model conforms to characteristics observed in Flickr and Yahoo!360, i.e. nodes in network are classified into three classes: inactive node with degree of zero (isolate node), huge branches with strong connection internal and all kinds of isolate small communities. Among them, isolate communities are basically in star topology structure, which expand fast at first and merge into a huge branch or stop expanding later. In huge branches, average distance between node decreases along with lapse of time. During network evolution process, new nodes and connections continuously appear. Increase of nodes directly leads to expansion of network scale while increase of edges may lead to the merger of two separated parts in the network. Along with the lapse of time, network scale changes and different branches merge, but the proportions of the three parts in the network basically remain the same.

2.6 Summary

In this chapter, we take social network structure as the research object. Based on important characteristics of social network, we detailed several important structure parameters of social network, i.e. degree distribution, average path length and so on, then carried out detailed analysis upon small-world phenomenon, scale-free characteristic, assortativity and other structural characteristics of social network, which are different from other complex networks. On this basis, we continued to introduce the establishment method of WS small-world model, BA scale-free model, forest-fire model, Kronecker graph model, production model and other traditional social network structure model. The principal purpose for establishing social network structure model is to research the generation mechanism of certain network structures and some network properties, and, in the second place, conveniently and economically provide simulation data from real network for researches in other fields. Structure characteristic research on social network is the basis of researches on other aspects and its continuous development significantly promoted social network science as well as complex network science.

Research and analysis on online social network is one of the fields first paid attention to and fully explored. During 2005 to 2010, structure characteristic work based on real social network data is active, covering online social networks home and abroad from familiar Flickr and Livejournal to uncommon Cyworld and Wealink, from foreign Twitter and Facebook to domestic Sina Weibo and Xiaonei. These works enable us to fully recognize similarity and characteristics of different online social networks, and obtain a series of parameters available for describing network structure characteristic. In recent years, research only focusing on network structure becomes rare along with applying results in earlier stages to other subsequent research fields, such as node impact prediction by network structure parameters by applying statistical method, time length prediction on information dissemination.

References

- [1] Mayhew Bruce, Levinger Roger. Size and the Density of Human Interaction in Social Aggregates[J].American Journal of Sociology, 1976, 82(1): 86-110.
- [2] Xiaofan Wang, Li Xiang, Chen Guangrong. Complex Network Theory and Its Application. Beijing:

Tsinghua University Press, 2006.

- [3] Mislove Alan, Marcon Massimiliano, Gummadi Krishna P, et al. Measurement and analysis of online social networks[C]. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007: 29-42.
- [4] Ahn Yong-Yeol, Han Seungyeop, Kwak Haewoon, et al. Analysis of topological characteristics of huge online social networking services[C]. Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 835-844.
- [5] Milgram Stanley. The small world problem[J]. Psychology today, 1967, 2(1): 60-67.
- [6] Backstrom Lars, Boldi Paolo, Rosa Marco, Ugander Johan, Vigna Sebastiano(2011-11-19). “Four Degrees of Separation”. ArXiv. Retrieved 23 November 2011.
- [7] Barabasi Albert-Laszlo, Albert Reka. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.
- [8] Erdős Paul, Rényi Alfred. On random graphs I[J]. Publ. Math. Debrecen, 1959, 6: 290-297.
- [9] Karl Sigman. Appendix: A primer on heavy-tailed distributions. Queueing Systems, 33(1):261-275, 1999.
- [10] Aaron Clauset, Cosma Shalizi, and Mark Newman. Power-law distributions in empirical data. SIAM Review, 2009, 51(4):661-703.
- [11] Kwak Haewoon, Lee Changhyun, Park Hosung, et al. What is Twitter, a social network or a news media?[C]. Proceedings of the 19th international conference on World wide web. ACM, 2010: 591-600.
- [12] Ravasz Erzsebet, Barabasi Albert-Laszlo. Hierarchical organization in complex networks[J]. Physical Review E, 2003, 67(2): 026112.
- [13] Wilson Christo, Boe Bryce, Sala Alessandra, et al. User interactions in social networks and their implications[C]. Proceedings of the 4th ACM European conference on Computer systems. Acm, 2009: 205-218.
- [14] Fu Feng, Chen Xiaojie, Liu Lianghuan, et al. Social dilemmas in an online social network: the structure and evolution of cooperation[J]. Physics Letters A, 2007, 371(1): 58-64.
- [15] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time:densification laws, shrinking diameters and possible explanations. In Proc. of ACM SIGKDD, pages 177-187, Chicago, Illinois, USA, 2005. ACM Press.
- [16] Feng Fu, Lianghuan Liu, and Long Wang. Empirical analysis of online social networks in the age of web 2.0. Physica A: Statistical Mechanics and its Applications, 2008, 387(2):675-684.
- [17] Haibo Hu and Xiaofan Wang. Evolution of a large online social network. Physics Letters A, 2009, 373(12):1105-1110.

- [18] Gong Neil Zhenqiang, Xu Wenchang, Huang Ling, et al. Evolution of social-attribute networks: measurements, modeling, and implications using google+[C]. Proceedings of the 2012 ACM conference on Internet measurement conference. ACM, 2012: 131-144.
- [19] Newman Mark, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.
- [20] Diego Garlaschelli and Maria I Loffredo. Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26):268701, 2004.
- [21] Cha Meeyoung, Mislove Alan, Gummadi Krishna P. A measurement-driven analysis of information propagation in the flickr social network[C]. Proceedings of the 18th international conference on World wide web. ACM, 2009: 721-730.
- [22] Holme Petter, Edling Christofer R, Liljeros Fredrik. Structure and time evolution of an Internet dating community[J]. *Social Networks*, 2004, 26(2): 155-174.
- [23] Yongjun Li. Topological Characteristics Analysis of Online Social Network [j]. *Complex System and Complexity Science*, 2012.
- [24] Watts Duncan James, Strogatz Steven Henry. Collective dynamics of ‘small-world’ networks[J]. *Nature*, 1998, 393(6684): 440-442.
- [25] Wang Xiao Fan, Chen Guanrong. Complex networks: small-world, scale-free and beyond[J]. *Circuits and Systems Magazine, IEEE*, 2003, 3(1): 6-20.
- [26] Newman Mark, Watts Duncan James. Renormalization group analysis of the small-world network model[J]. *Physics Letters A*, 1999, 263(4): 341-346
- [27] Newman Mark. The structure and function of complex networks[J]. *SIAM review*, 2003, 45(2): 167-256.
- [28] Boccaletti Stefano, Latora Vito, Moreno Yamir, et al. Complex networks: Structure and dynamics[J]. *Physics reports*, 2006, 424(4): 175-308.
- [29] Cohen Reuven, Havlin Shlomo. Scale-free networks are ultrasmall[J]. *Physical Review Letters*, 2003, 90(5): 058701.
- [30] Fronczak Agata, Fronczak Piotr, Holyst Janusz. Mean-field theory for clustering coefficients in Barabasi-Albert networks[J]. *arXiv preprint cond-mat/0306255*, 2003.
- [31] Albert Reka, Barabasi Albert-Laszlo. Topology of evolving networks: local events and universality[J]. *Physical review letters*, 2000, 85(24): 5234.
- [32] Bianconi Ginestra, Barabasi Albert-Laszlo. Bose-Einstein condensation in complex networks[J]. *Physical Review Letters*, 2001, 86(24): 5632.
- [33] Bu Tian, Towsley Don. On distinguishing between Internet power law topology generators[C]. *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications*

- Societies. Proceedings. IEEE. IEEE, 2002, 2: 638-647.
- [34] Chakrabarti Deepayan, Zhan Yiping, Faloutsos Christos. R-MAT: A Recursive Model for Graph Mining[C]. SDM. 2004, 4: 442-446.
- [35] Leskovec Jure, Faloutsos Christos. Scalable modeling of real graphs using kronecker multiplication[C]. Proceedings of the 24th international conference on Machine learning. ACM, 2007: 497-504.
- [36] Leskovec Jure, Chakrabarti Deepayan, Kleinberg John, et al. Kronecker graphs: An approach to modeling networks[J]. The Journal of Machine Learning Research, 2010, 11: 985-1042.
- [37] Kumar Ravi, Novak Jasmine, Tomkins Andrew. Structure and evolution of online social networks[C]. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 611-617.
- [38] Leskovec Jure, Horvitz Eric. Planetary-scale views on a large instant-messaging network [C]. Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 915-924.
- [39] Golbeck Jennifer. Analyzing the social web[M]. Newnes, 2013.

Technologies and Approaches for Virtual Community Detection

3.1 Introduction

Along with the development of communications and computer technology, social network now serves as the important platform for daily dating and communication, personal life show and message-distributing. The relationships in social network users are uneven. Some individuals have dense relationships while others have sparse relationships, which form the virtual community structure in social network. Detecting the community structure in social network contributes to understanding the characteristics of topological structure of network, revealing the intrinsic characteristics of complex systems and comprehending the relationships/behaviors of individuals as well as its evolution trend, thus providing strong support for information retrieval, information recommendation, information propagation control, organization management, public safety incidents control and many other applications. Virtual community detection in social network has both important theoretical value and great practical significance. In recent years, research on algorithms for virtual community detection has attracted much attention of scholars. They have put forward a series of classic community detection algorithms for mining virtual communities in social networks of different sizes. What is more, along with the increase of social network size and node information, community detection algorithms are trying to achieve high accuracy while reducing the time complexity and pay more attention to using the local topological structure of social networks.

This chapter will introduce details related to virtual community detection algorithms, including evaluation system of the algorithm and some classic algorithms for community detection. The content of this chapter is organized as follows: First, Section 3.2 presents the definition of virtual community and development process of community detection algorithms, and introduces the evaluation system of the algorithm in two aspects of evaluation accuracy and computation complexity. Second, according to the different computation processes of objective functions in the community detection algorithms, we classify the algorithms into two classes and give descriptions respectively. Section 3.3, from the perspective of the static computation, introduces some classic static community detection algorithms such as modularity optimization algorithms, multi-objective optimization algorithms, algorithms based on probability model and information coding algorithms. Section 3.4, from the perspective of dynamic computation, introduces some classic dynamic algorithms such as cluster percolation algorithms, agglomerative algorithms based on similarity, label propagation algorithms and local expansion optimization algorithms.

3.2 Theoretical Basis of Virtual Community Detection Technology

3.2.1 The Definition of Virtual Community

In the 1970s, many scholars began to realize that there are some node sets with closely-connected nodes in many graphs. These sets have a great influence on topological structure of the entire graph. As a result, people started to use mathematical tools such as graph theory to describe the detection of those node sets as the problem of graph partitioning and defined those sets as subgraph with the characteristic of close connection. Along with the gradual improvement and deepening of the complex network theory, people realized node sets with the characteristic of locally close connection also exist in complex networks and scholars put forward the concept of community. Mark Newman and other complex system scientists tried to reveal and expound such community structure by some theories of complex networks. In recent years, online social network has progressed vigorously as a new type of complex network, in which nodes in the virtual network are mapping of people in the real society and network edges denote the exchange and

communication between network users. At present, the problem of community detection in online social networks has attracted much attention. As online social networks can be regarded as virtual environment platform different from the real world, community can also be called as virtual community. So far, people have given many different definitions of the community (virtual community) from different aspects, including the local definition based on subgraph, the global definition based on the network modularity, definition based on the similarity between nodes and other typical definitions.

1. Local Definition Based on Subgraph

Community structure can be regarded as several node sets with high cohesion in the network topology. These sets are usually abstraction of relatively independent components with independent functions or properties. Therefore, the community structure can be defined according to the characteristics of local network topology. So far, a descriptive definition accepted by scholars in various fields is the local definition based on subgraph, that is, community structures are several subsets of the node sets in complex network. The connections between nodes in each subset are relatively very compact, while the connected edges between nodes in different subsets are relatively sparse. In Figure 3-1, 20 nodes in the network are partitioned into three community structures, and each one corresponds to the structure in three dotted circles. In these three communities, nodes are closely connected to each other, while edges between communities are relatively sparse.

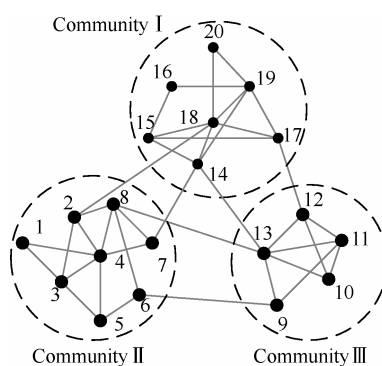


Figure 3-1 A small network with virtual community structure

2. Global definition based on the network modularity

In order to evaluate the community structure, Mark Newman et al. put forward the

definition of modularity^[1] by simulating the definition of variance according to the differences between topological structures of real networks and random networks. The modularity can be defined as the expectation of the difference between the ratio of edges between nodes in the same class (edges in the same community) and the ratio of edges between these nodes of random connection in the same community structure. According to this definition, the higher modularity brings better community structure. The modularity can evaluate the community structure through the global information in topological structures of networks. This community definition based on modularity is often applied in community detection algorithms to evaluate the community structure or provide stop condition for the algorithm.

3. Definition Based on the Similarity Between Nodes

In physical sense, communities usually denote set of elements in complex systems or networks with the same or similar function. These elements collaborate or interact with each other, completing some relatively independent functions in the entire system or forming some relatively independent organizational structures. As a result, communities can be defined based on the similarities between nodes. This definition assumes that nodes in the community are similar, while the similarity between nodes in different communities is low. Certain index should be used to evaluate the similarity between network nodes for further definition of community structure.

In general, from the aspect of essential meaning, all the existing definitions of community (virtual community) are consistent, that is, for a subset of the set comprised by all individuals in networks, individuals therein are closely connected based on certain property and have sparse connection with individuals out of the subset.

3.2.2 Development Process of Virtual Community Detection Algorithms

Essentially, virtual community detection in online social networks can be regarded as the procedure of partitioning the network nodes into several subgraphs according to the closeness in topological structure. In computer science, such problems are often regarded as graph partitioning problem. Research on graph partitioning problem dates back to 20th century, among which two most important algorithms are Kernighan-Lin algorithm^[2] and the spectral bisection algorithm^[3]. Kernighan-Lin algorithm, based on greedy optimization

strategy, defines a objective revenue function and seeks for the best partitioning through which the objective revenue function can achieve maximum value by way of greedy search. Kernighan-Lin can not only find the reasonable partitioning of network, but also show the community structure by tree diagram, thus revealing the hierarchical community structure. The spectral bisection algorithm starts from the Laplacian matrix of network and implement dichotomy on the network topological structure by researching the eigenvalue and eigenvector of the matrix. Community partition with the number of community greater than 2 can be obtained by applying the algorithm iteratively. In the 21st century, along with the development of complex network science, the problem of network community structure detection has attracted more attention of experts from different fields. Michelle Girvan and Mark Newman put forward a new splitting algorithm in 2002, i.e. the GN algorithm^[4]. In the algorithm, they put forward the concept of modularity to evaluate the community structure based on the comparison between structural characteristics of complex networks and random networks, initiating prosperous development of community detection. GN algorithm, essentially a kind of network splitting algorithm, identifies edges between communities through customized edge betweenness and removes the edge with biggest edge betweenness, thus splitting the network into several virtual communities. Furthermore, Mark Newman et al. found that optimizing the objective function modularity helps to better detect community structure, as modularity is an important index to evaluate the community partition. Inspired by this idea, many scholars take the modularity as the objective function and put forward many community detection algorithms^[20] based on the optimization of modularity function. Considering essential deficiency in the modularity function, Shi Chuan et al. put forward the community detection algorithm based on multi-objective optimization. They described the community structure characteristics accurately and comprehensively with multi-objectives function and realized optimization of objective function to detect the community structure in network. For applications in need of detailed description of community structure, experts in the field of information science, from the aspect of information theory, mapped the network topological structure into data coding problem and achieved community detection by constructing the community partition with shortest coding length. A typical example is Infomap algorithm, which can detect the communities in networks more accurately and is often applied in the occasion in need of accurate analysis for community structure^[5]. Considering the phenomenon that several nodes in networks belonging to multiple communities at the same time, Gergely Palla et al. put forward the concept of overlapping community in 2005 and tried to detect the

overlapping communities in networks and bridge nodes along the boundary of communities based on the definition of clique^[6]. By researching on the differences in topology characteristics between real networks and virtual networks through mathematical tools such as Bayesian inference, Mark Newman et al. put forward the community detection algorithm based on probability model and detected overlapping communities by the maximization of the likelihood probability. All the algorithms above put more emphasis on the accuracy of community structure while pay less attention to the time complexity, thus the time complexity is often relatively high. As a result, these algorithms are only fit for virtual community detection in social networks with small size. In fact, it is often necessary to analyze the virtual communities in social networks with large-scale and wide range of node information, requiring the community detection algorithm to have high accuracy and low time complexity. Considering requirements from both aspects, scholars put forward some new community detection algorithms from different perspectives for accurately and rapidly detection of virtual communities in social networks with large-scale. People researched on local structure characteristics of community structures and put forward community detection algorithms based on local expansion. This kind of algorithms often start from one or several core nodes in the network, define the local objective function and absorb surrounding nodes into existing virtual community structure by greedy strategy. As this kind of algorithms only uses the local topological structure of network, it is capable of detecting and analyzing the community structure in the area concerned before constructing topological structure diagram of the entire network^[7]. To improve the efficiency of community detection in networks, Usha Nandini Raghavan et al. put forward a community detection algorithm based on label propagation^[8]. The algorithm provides a unique label for each node, and iteratively updates the label value of a node according to the majority labels of its neighbors. Finally, the procedure converges into the stable status with some nodes sharing the same label in the network, which are nodes belong to the same community. This kind of algorithms can detect the network community structure in linear time when the topological structure of the entire network is known.

3.2.3 The Accuracy Indexes of Evaluation for Virtual Community Detection Algorithms

So far, various kinds of algorithms have been proposed for community detection. Different algorithms may partition the same network into different community structures. It

is a big challenge to evaluate those community structures partitioned by different algorithms. For this purpose, people put forward some digital evaluation indexes^[9] to measure the accuracy of community detection algorithms, such as modularity and NMI.

1. Modularity

Mark Newman put forward the modularity index to measure the differences in community structure by comparing the connection density differences between existing networks and reference networks under the same community partitioning, with reference network is the random network with the same degree sequence as original network. Assume that A denotes the adjacent matrix of a complex network and k_v denotes the degree of node v , i.e. $k_v = \sum_w A_{vw}$. In corresponding reference network, the probability of existence

of edge (v, w) is $\frac{k_v k_w}{2m}$, with m as the number of edges in network graph A . Then the complete mathematical expression of modularity is as shown in Formula (3-1):

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w) \quad (3-1)$$

where c_v denotes the community which v belongs to. If $i = j$, $\delta(i, j) = 1$; otherwise $\delta(i, j) = 0$. In mathematical sense, Formula (3-1) denotes the expectation difference between the ratio of edges in the same community and the ratio of edges in reference network under the same community structure. Higher modularity brings better community partitioning in complex networks. In order to calculate the modularity Q more conveniently, use the following Formula (3-2). Assume that the complex network is partitioned into k community structures and define a symmetric matrix $e(e_{ij})$ with $k \times k$ order, where e_{ij} denotes the ratio of edges between two communities i and j to all edges of the network. The sum of all elements on the diagonal denotes the ratio of edges between nodes in the same community to all edges of the entire network, expressed by $\text{Tre} = \sum_i e_{ii}$. And

the sum of elements in each row is expressed by $a_i = \sum_j e_{ij}$, which denotes the ratio of edges connected to nodes in community i to edges of the entire network.

Based on the definition above, Formula (3-1) can be transformed into Formula (3-2).

$$Q = \sum_i (e_{ii} - a_i^2) = \text{Tr} - ||e^2|| \quad (3-2)$$

where $||e||$ denotes the sum of all elements in matrix e . More obvious community structure in complex networks results in bigger modularity Q . In real networks, the value range of modularity is often 0.3~0.7.

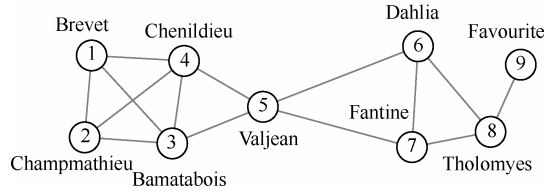


Figure 3-2 Relationships between some characters in Les Misérables

Taking Figure 3-2 for example, the computation process can be introduced as follows. The figure is an abstraction reflecting the relationships between some characters in Les Misérables, where node 1, node 2, node 3 and node 4 denote the characters related to the Champmathieu case, while node 6, node 7, node 8 and node 9 denote characters centered by Fantine, mother of the daughter adopted by Valjean. The adjacent matrix corresponding to the figure is expressed as follows:

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Assume that, through a certain algorithm, the network can be partitioned into three communities: $c_1=\{1, 2, 3, 4\}$, $c_2=\{5\}$, $c_3=\{6, 7, 8, 9\}$. $m=14$ denotes the number of edges in the figure. The modularity component of community c_1 is 0.1786, the modularity component of community c_2 is 0 and the modularity component of community c_3 is 0.1581, with total modularity degree of 0.3367.

Although the modularity proposed by Mark Newman can evaluate the community partitioning accurately, as recognized by many specialists and scholars, there are still

several problems in modularity indexes. For example, Santo Fortunato et al. found that the partitioning corresponding to the modularity maximum is not necessarily the best community partitioning result. Under many circumstances, there is a potential scale of the smallest community, and any community structure with scale smaller than such potential will cause negative effect on the modularity optimization.

2. NMI

Along with the development of online social networks, people realized that lots of online social networks have information implying the community membership of each node. For example, the school information on renren.com reveals the community structure of nodes from the same school and the interest information on Facebook characterize the virtual user groups with same interests. These data provides rich information for virtual community detection as well as the standard answer for evaluating virtual community structure. In case that some virtual community structure information is known in advance, Leon Danon et al. put forward the Normalized Mutual Information (NMI), which evaluate the differences between the community structure partitioned by the algorithm and the known community structure^[10]. NMI is a digital index calculated by confusion matrix N . Given two community partitions: $a = (a_1, a_2, \dots, a_n)$, $b = (b_1, b_2, \dots, b_n)$, where a_p , b_p ($p = 1, 2, \dots, n$) denote the number of community respectively which node p belongs to in these two partitions and n denote the number of nodes in the network. NMI formula is as follows:

$$\text{NMI} = \frac{-2 \sum_{i,j} N_{ij} \ln \left(\frac{N_{ij} n}{N_{i.} N_{.j}} \right)}{\sum_i N_{i.} \ln \left(\frac{N_{i.}}{n} \right) + \sum_j N_{.j} \ln \left(\frac{N_{.j}}{n} \right)} \quad (3-3)$$

where $N_{i.}$ denotes the sum of elements in row i of the matrix N and $N_{.j}$ denotes the sum of elements in column j of matrix N .

This numeric index can be used to evaluate the difference between detected community structure and known structure. Higher value results in better partitioning of community structure. If the value reaches the maximum 1, the community structure detected by the algorithm is the same as known structure and the result of the algorithm is the best.

With the example of Figure 3-2, the computation procedure of NMI is shown as follows. Assume that the best known community structure partitioning can be expressed as

$\{1, 2, 3, 4\}$, $\{5\}$ and $\{6, 7, 8, 9\}$. The vector of community partitioning is $\mathbf{a} = (1, 1, 1, 1, 2, 3, 3, 3, 3)$. Then assume that the community structure partitioned by a certain algorithm can be denoted by vector $\mathbf{b} = (3, 3, 3, 3, 2, 1, 1, 1, 1)$.

According to the known community partition vector, the confusion matrix N can be constructed as follows:

$$N = \begin{pmatrix} 0 & 0 & 4 \\ 0 & 1 & 0 \\ 4 & 0 & 0 \end{pmatrix}$$

According to Formula (3-3), the NMI value of this partitioning is 1.

3. Rand Index

Virtual community detection problems in online social networks can be regarded as the problem of clustering in the field of data mining. Except for NMI numeric index, scholars in data mining field put forward some other numeric indexes to evaluate the clustering results, among which the Rand index is a typical example^[11]. Rand index indicates the ratio of number of node pairs which belong to the same community or different communities in both two partitions. Assume that n nodes are denoted by X_1, X_2, \dots, X_n in two partitioning: Y and Y' . The Rand index can be calculated as follows:

$$R(Y, Y') = \sum_{i < j} \gamma_{ij} / C_n^2 \quad (3-4)$$

where C_n^2 denotes the number of probable node pairs of the n nodes, i.e. $(n-1)n/2$. If node X_i and X_j are partitioned into the same community or different communities in both two partitioning, $\gamma_{ij} = 1$; otherwise, if the node X_i and X_j are partitioned into the same class in one partitioning but partitioned into different communities in another partitioning, $\gamma_{ij} = 0$.

The computation procedure can also be illustrated by Figure 3-2. Assume that there are two partitioning $\{\{1, 2, 3, 4\}, \{5\}, 6, 7, 8, 9\}\}$ and $\{\{1, 2, 3, 4, 5\}, \{6, 7, 8, 9\}\}$. According to the definition of γ_{ij} , 12 node pairs are partitioned into the same class in both partitioning and 20 pairs are partitioned into different classes in both two partitioning. There are 36 node pairs in total, so the Rand index of these two partitioning is $8/9$. If these two partitioning are the same, the value is 1 and results of the algorithm is the best. For its simple calculation in definition, this index is capable of effectively evaluating the partition

results of virtual communities in online social networks of large scale.

An improvement for Rand index is Mirkin with the formula as follows:

$$M(X, Y) = n(n-1)[1 - R(X, Y)] \quad (3-5)$$

4. Jaccard Index

Jaccard index is a numeric index similar to Rand index with the formula as follows:

$$J(x, y) = \frac{a_{11}}{a_{11} + a_{01} + a_{10}} \quad (3-6)$$

where a_{11} denotes the number of pairs partitioned into the same subset in both partitioning while a_{01} and a_{10} denote the number of pairs partitioned into the same community in one partitioning but partitioned into different communities in another partitioning. We introduce the computation process of Jaccard index with the example in Figure 3-2. Considering two kinds of partitioning: $\{\{1, 2, 3, 4\}, \{5\}, \{6, 7, 8, 9\}\}$ and $\{\{1, 2, 3, 4, 5\}, \{6, 7, 8, 9\}\}$ with $a_{11} = 12$ and $a_{11} + a_{01} + a_{10} = 16$, then $J(x, y) = 12/16 = 0.75$.

3.2.4 The Computational Complexity of Algorithms for Virtual Community Detection

To evaluate a community detection algorithm, we should consider its computational complexity in addition to its accuracy in partitioning. In terms of the computational complexity of algorithms, the amount of resources required for the task mainly depends on the time complexity and space complexity.

In general, time complexity of an algorithm is the function of computation effort for executing the algorithm, which describes run time of the algorithm quantitatively. Generally, we use the symbol O to denote the function, which provides the maximum of the run time. In this way, the expression denotes the time magnitude needed for the algorithm when the input tends to infinite, without any lower order terms and leading coefficients.

Space complexity measures temporary memory space needed for the algorithm at run time, which is usually the function of problem scale. In the field of virtual community detection, we usually denote problem scale as the number of nodes or edges in the network. Similar to time complexity, we use O to denote the scale function of space complexity for the problem in space complexity.

3.2.5 Typical Data Sets Needed for Testing Virtual Community Detection Algorithms

To test the performance of the community detection algorithms, many scholars in various fields, especially sociology, have carried out modeling and abstraction for complex systems in different fields, and extracted many topological reference graphs with typical community structure. Centered on the detection problem of community structure, typical datasets can be classified into two classes: real reference network and artificial reference network.

1. Real Reference Network

Real reference network is an abstraction of real social networks with obvious community structure, thus the community structure therein usually has specific practical meaning. The network of Zachary's karate club is most cited in the research area of current community detection algorithms for social network, which reflects the social relationship between members in the karate club^[12]. In the early 1970s, Zachary observed the relationship between members of a karate club in a college in America for two years. In this graph, Zachary club members are denoted by nodes and social relationships by edges between nodes in the network, thereby constructing the relationship network of the club can be constructed as shown in Figure 3-3. In the network, 34 nodes denote 34 members and edges between nodes denote friendship between them. Within two years after Zachary constructs this network graph, the club administrator (node 1) and the instructor (node 33) broke up on whether to raise the fee of the club, and the club split into two communities centered on the club administrator and the instructor respectively. These two communities provide good results evaluation basis for partitioning virtual community in social network and cited by many scholars. However, this network has only 34 nodes, which is its main deficiency. In the 21st century, scholars from various fields tried to find a network with larger scale and thereby detect virtual community structure. Mark Newman et al. analyzed and sorted the schedule arrangements between teams of MLS, thereby extracted the NCAA football network with the topological structure of the network as shown in Figure 3-4. In the NCAA football network, there are 115 nodes denoting the teams and each edge denotes one or several matches played between two teams. These 115 teams are partitioned into 12 communities by the states they belong to with more matches played between teams in each community and less matches played between teams between communities, which form a natural community structure.

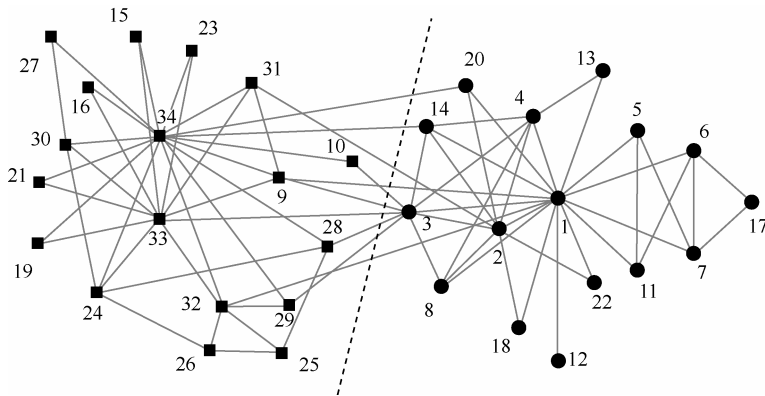


Figure 3-3 Topological structure of Zachary's karate club network

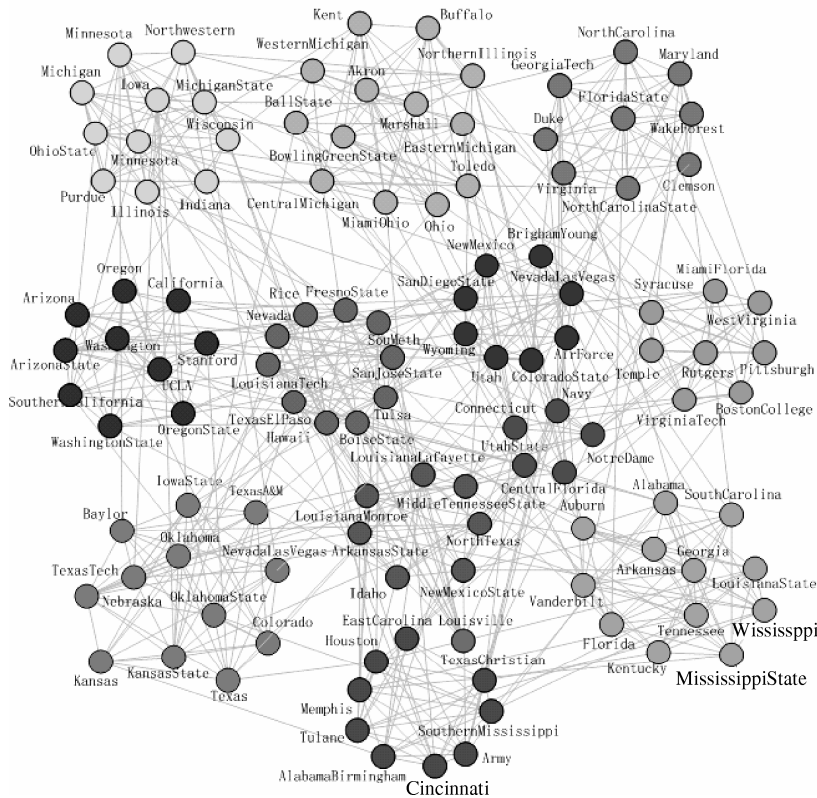


Figure 3-4 Topological structure of NCAA MLS

The development of online social networks provides lots of data for research on virtual community detection algorithms. Scholars realized that many characteristics in

online social networks reflect the social relationship in real world with many special topological properties. These properties have great influence on the virtual community structure of online social networks, through which many scholars collect the information in many online social networks as the practical test data for virtual communities. The data set of blog network of U.S. politicians is a typical example. The data set includes 1,491 nodes and each node denotes a virtual ID on a famous blog. During the U.S. presidential election 2004, users on the blog are partitioned into conservatives and liberals according to their different political stands, leading to the obvious communities in the network. Figure 3-5 shows the community structure formed by part of nodes in the network which are partitioned into two communities according to their different political preferences and marked in black and grey respectively. The data set collected from a famous online social site in U.S. intuitively reflects the essential features of community structure in the network. The Blog network based on MSN, an online chatting tool, is another typical example. Each user on MSN is denoted by a node in the network. The edge between two nodes can be constructed if the comment behavior between them is frequent. After several times of data collecting, the Blog network includes 30,557 nodes and 82,301 edges. In this network, account nodes with same interests have more mutual comments. As a result, the network reveals the virtual community structure marked by topics of interest. Along with the rapid development of online social networks like Facebook, many scientists collect the data therein and analyze the virtual community structure of such networks. Amanda Traud et al. construct the social network among students in a university in U.S. according to the similar relation of users^[41]. The network includes 769 nodes and 16,656 edges which are partitioned into several virtual community structures according to the friendship among users.

The relationship network of characters in *Les Misérables* introduced in Section 3.2.3 in this chapter is also a simple social network. As shown in Figure 3-6, 9 nodes in the network are partitioned into two communities. One community, based on Champmathieu case, is related characters interrogated instead of him after Champmathieu was mistaken for Valjean. The other community centers on Fantine which includes the social relationships among Fantine and people who abandoned her. Valjean, as the prolabelonist, is related to all the nodes in the network and can be regarded as the node in the overlapping region.

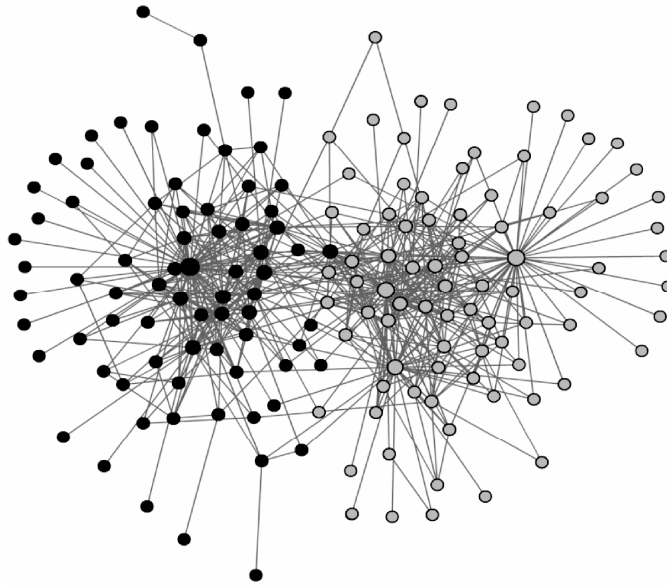


Figure 3-5 Blog network of U.S. politicians

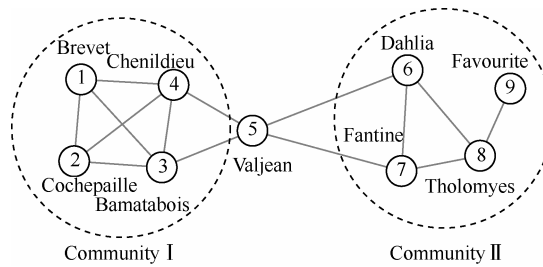


Figure 3-6 Network of character relationships in Les Misérables

2. Artificial Reference Network

Real reference networks are abstraction of systems in real world, and communities therein usually have certain background meaning. Systems are often affected by various factors in real world, so the community structure therein is not completely conform to the definition of community. To solve this problem, many experts and scholars in complex science and sociology constructed artificial network reference graph model based on the power-law distribution characteristic of node degree and the small-world phenomenon in complex networks. Without the influence of external factors in real world, the community structure in artificial network is more distinct and reasonable, providing a powerful means for evaluating community detection algorithms. Two famous network models in artificial

reference network are GN artificial network proposed by Mark Newman et al. and LFR artificial network proposed by Andrea Lancichinetti et al.^[13]. These two artificial networks both dynamically generate networks in specified node scale and with typical community structure based on the topological characteristics of complex networks with the number of nodes as parameter. As a result, they are suitable for constructing online social networks of large scale. GN artificial network, proposed by Mark Newman partition nodes in the network partitioned into different communities with the number of nodes in the network, number of communities and the edge distribution of each node as its input parameters. Degrees of each node, the number of edges inside the same community and other information decide the community that a node belongs to and the clarity of the entire network community structure. Figure 3-7 is a network topological structure graph with 128 nodes generated from GN artificial network, which is partitioned into 4 communities with 32 nodes in each community and each node therein has the degree 16.

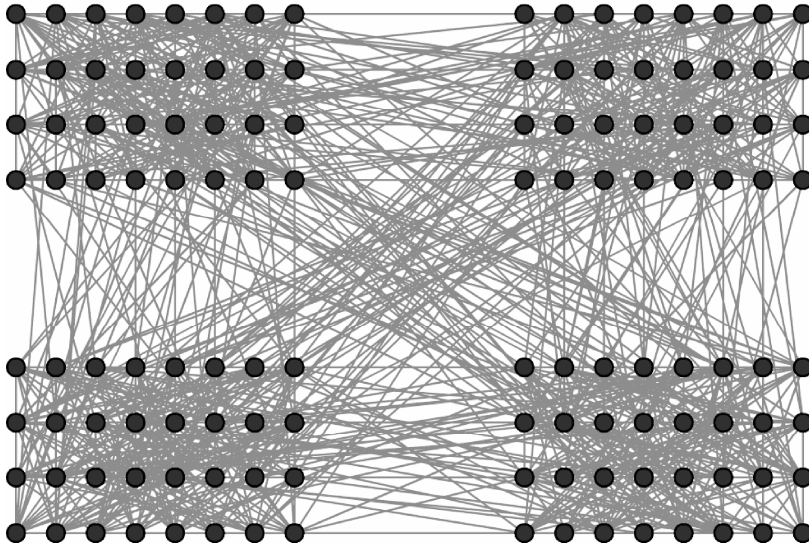


Figure 3-7 GN artificial network with 128 nodes

Although GN artificial network can simulate networks with community structure to a certain extent, each community has the same number of nodes because of the uniform partition of nodes into specified communities. This deficiency leads to big difference in topological properties between network generated from GN artificial networks and real complex networks. To address this problem, Andrea Lancichinetti et al. put forward the LFR reference network model. Compared with GN artificial network, the LFR artificial network needs more input parameters and better conforms to the real online social networks in topological properties

thanks to better flexibility of topological structure reference network constructed by it. In LFR artificial network, users can set parameters to control the entire network and topological properties of each community, such as community scale, degree distribution of nodes and the ratio between the number of nodes and number of edges in the same community. Figure 3-8 shows an artificial network with 5 communities generated by LFR. Compared with GN artificial network, in this network, number of nodes in each community is different and the node degree follows power-law distribution; therefore, it better conforms to topological structure characteristics of real networks. To simulate the characteristics of overlapping community structure in real networks, users can also set the number of overlapping nodes in LFR artificial networks to provide communities with overlapping structure for detection algorithms. Based on the features above, many scholars use LFR artificial network to simulate online social networks of large scale and thereby evaluate various property indexes of virtual community detection algorithms in online social networks.

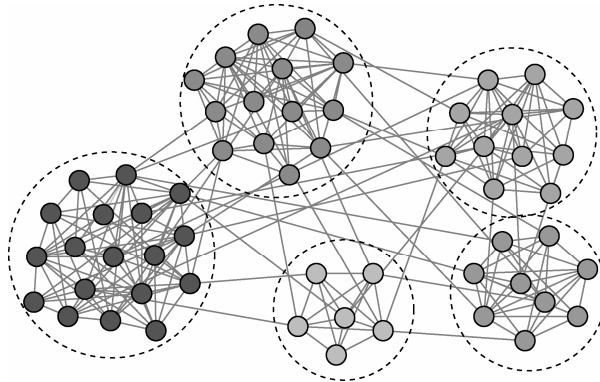


Figure 3-8 LFR artificial networks with 5 communities

3.3 Static Calculation Detection Algorithms for Virtual Communities

The existing network community detection algorithms can be classified into many classes based on different standards. For example, overlapping community detection algorithms and non-overlapping community detection algorithms based on whether the detected communities are overlapping or not; network topology-based algorithms, network dynamics-based algorithms, modularity function optimization-based algorithms and other algorithms based on different physical background; static calculation algorithms and dynamic calculation algorithms based on

different calculation mechanism. For static calculation, each calculation step considers all nodes in the network and calculates whether certain partitioning conforms to global optimization objectives, thus, determines the final community structure. For dynamic calculation, starting from local nodes, it update local node status according to certain rules and gradually infer the final global partitioning results of all nodes, while intermediary steps therein don't need to meet the global optimization objective of all nodes. Static calculation detection algorithms are introduced in this section and dynamic calculation algorithms will be introduced in the next section.

3.3.1 Modularity Optimization Algorithms

As mentioned above, Mark Newman put forward the concept of modularity to measure the intensity of community structure and is defined as (see Section 3.2.3)

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j) \quad (3-7)$$

The modularity value mainly depends on community partitioning of nodes in network C , i.e. the community partitioning situation in network, which can quantitatively measure the quality of community partitioning. Higher modularity value indicates better partitioning. Thus, optimal community partitioning of networks can be achieved by maximizing the modularity Q . The number of possible partitioning of a network is enormous. If the number of nodes and edges of a network are respectively denoted as n and m , the number of possible partitioning is exponent of n . Thus it is a NP-hard problem to find out optimal partition among all possible partitions. Some algorithms have been proposed to detect approximate optimal partition maximizing the modularity in reasonable time. Some of them are introduced below.

1. Classic Greedy Algorithm

Mark Newman put forward a greedy modularity optimization algorithm FN^[1]. Greedy algorithm aims to find overall optimal value or approximate optimal value of objective function. It decomposes entire optimization problem into multiple local optimization problems and find out optimal values of all local optimization problems which are integrated into overall approximate optimal value. In this sense, FN decomposes modularity optimization problem into local modularity optimization problems. First, regard each node of network as a small independent community; second, calculate the modularity gain of merging each two connected communities. Two communities whose combination leads to

largest modularity gain or smallest modularity loss are selected to merging into a new community according to greedy principle. Repeat the iteration until all nodes are grouped into one community. The modularity values are changing along with the iteration and the community partition corresponding to largest modularity value is regarded as the approximate optimal one.

The specific steps of greedy algorithm FN are described below.

(1) Remove all edges in the network and regard each node of a network as an independent community.

(2) Regard each connected part as a community in the network and add edges out of the network back to the network, one edge at a time. If the added edge connects two different communities, merge these two communities and calculate the modularity increment of the newly-formed community. Choose the two connected communities resulting in the greatest increase (or smallest decrease) in modularity for merging.

(3) If the number of communities is larger than 1, return to step (2), otherwise go to step (4).

(4) Select the partition with largest modularity among all iterations as the optimal community partition for the network.

In this algorithm, it should be noted that newly-added edge only affect the community partitioning of the network, and each calculation of modularity of network partitioning is completed on the intact topological structure of the network, i.e. the network including all edges in the network.

The FN algorithm is further illustrated based on Figure 3-2 below for better understanding of this algorithm,.

(1) Initially, Remove all edges in the network and regard each node as an independent community, 9 communities totally.

(2) Add edges out of the network back to the network and calculate the modularity increment of the new community partition. The largest increase of modularity is obtained when merging {8, 9} at $\Delta Q = 0.064$. Thus community {8} and {9} are firstly merged into community {8, 9}.

(3) Repeat step (2) for updated 8 communities until all nodes are grouped into one community. There are 9 different community partitions totally along with the iterations with the tree diagram of communities partitioned by FN shown in Figure 3-9.

(4) The community structure with largest modularity value is obtained when network is partitioned by dash line in Figure 3-9. The largest modularity value is $Q = 0.36$ and the

resulted community structure contains two communities, i.e. $\{1,2,3,4\}$ and $\{5,6,7,8,9\}$.

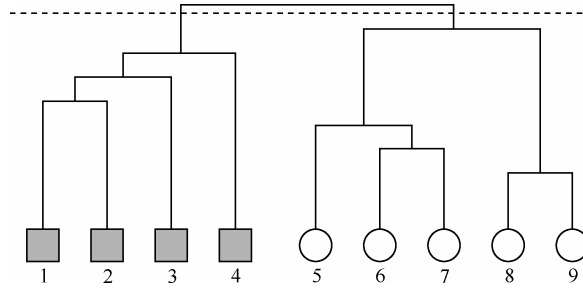


Figure 3-9 Dendrogram of communities partitioned by FN algorithm on example network

The calculation performance of this algorithm is as follows. Assume that the network has n nodes and m edges. There are $n-1$ iterations in FN. Since one pair of communities is merged in each iteration, $n-1$ join operations are carried out totally to construct the complete dendrogram. In each join, at most m possible pairs of communities need to be checked in time $O(m)$, and joining two communities takes $O(n)$ to update the adjacency matrix of network. Thus, the total time complexity of FN is $O((m+n)n)$, which is approximate to $O(n^2)$ in sparse networks. Though Mark Newman uses this algorithm to analyze a co-authorship network with 56,276 nodes successfully, the time complexity of FN is still large. FN is only suitable to detect community structures in small social networks.

There are other heuristic algorithms for obtaining optimal modularity like simulated annealing algorithm^[15], extremum optimization algorithm^[16], etc. Simulated annealing algorithm obtains optimal modularity value more close to real maximum but requires larger time complexity. Extremum optimization algorithm obtains slightly worse modularity but much better time complexity.

2. Fast Modularity Optimization Algorithm

In order to reduce time complexity, Vincent Blondel et al. proposed a hierarchical greedy algorithm^[17] which includes two labels. In first stage, regard each node as a community and decide the neighbor communities to be merged based on maximized modularity increment. The second stage begins after a round of scanning. All communities resulting from first stage are regarded as nodes to construct a new network. Repeat the first stage in such new network. Repeat the two labels alternatively until the modularity value ceases to increase and obtain the approximate optimal community partition of a network.

This algorithm has some advantages. First, the algorithm is intuitive and easy to implement. Secondly, the number of communities does not need to be set in advance. Thirdly, it can present the hierarchical virtual community structures of an online social network and detect community structures in different resolutions. Finally, simulation experiments show that the algorithm has nearly linear time complexity in sparse networks. It can partition the network with more than 10^9 nodes in reasonable time. Thus this algorithm is suitable for detecting community structures in online social networks, a kind of complex network with super-large-scale.

3.3.2 Multi-objective Optimization Algorithms

Optimization algorithms represented by modularity optimization algorithm convert the virtual community detection problem to the extremum optimization problem, thereby solve the problem by heuristic algorithms such as simulated annealing algorithms. Online social networks are abstraction of real social networks, in which communities usually have several kinds of characteristics in structures and properties and are difficult to be described by a single characteristic. To solve this problem, many scholars detect virtual community structures in online social networks by adopting multi-objective optimization theory based on traditional modularity optimization algorithms. Several multi-objective optimization algorithms are given below.

1. Community Detection Algorithms Based on Cellular Automata

As mentioned above, traditional community detection algorithms based on modularity have many defects. For example, recent research showed that a community would be partitioned into large adjacent community through community detection algorithms based on modularity when the scale of community structure is too small. In this case, the value of modularity is big but the partition may be unreasonable. Moreover, communities in online social networks usually have several kinds of characteristics in structure and properties and are difficult to be described by a single characteristic. To solve this problem, many scholars described the community structure in different aspects by adopting multi-objective optimization theory and partitioned virtual community structures by heuristic algorithms such as cellular automata principle and genetic algorithms.

Yuxin Zhao proposed a CLA-net algorithm based on cellular learning automata in which each node is regarded as a learning automata based on irregular cellular automata^[18]. Each

learning automata describes the community structure in aspects of community structure of the entire network and local community structure of nodes. Irregular cellular automata refers to the biological breeding phenomenon and creates a local dynamic model with discrete time dimension and space dimension. This model does not involve in compliance with strict mathematical equations or function, but develops some simple rules by repeated computation based on some definition of changing rules of cellular states to generate an extremely complex dynamic model. Yuxin Zhao et al. mapped the online social network as irregular cellular learning automata and adjusted the state of each node dynamically to make the community structure more reasonable by customized evolution rules.

To introduce the algorithm more specifically, some variants of learning automata L_i for each node i are defined as follows: α_i denotes the behavior aggregate of learning automata L_i and each alternative behavior corresponds to the sequence number of an adjacent node of node i ; p_i denotes the behavior probability vector of learning automata L_i and p_{ij} denotes the probability of behavior j ; $\alpha_i(t)$ denotes the behavior that L_i selects in the t^{th} iteration; $\beta_i(t)$ is the feedback signal accepted by L_i in the t^{th} iteration and $\beta_i(t)=0$ denotes the reward signal while $\beta_i(t)=1$ denotes the punishment signal. $W_{ij}(t)$ denotes times of the reward behavior j in the t^{th} iteration; $Z_{ij}(t)$ denotes times of behavior j in the t^{th} iteration; Q_{best} denotes the optimum value of modularity in current community structure.

The learning and updating process of learning automata L_i can be described as below:

- (1) Choose a behavior $\alpha_i(t)$ randomly according to the behavior probability vector p_i .
- (2) Interact with local environment (other adjacent nodes) and overall environment (the entire network), then obtain the feedback signal $\beta_i(t)$. If node i belongs to the same community as its most adjacent nodes and the modularity obtained in this iteration satisfies the inequality $Q(t) \geq Q_{best}$. Then the feedback signal $\beta_i(t)=0$, otherwise $\beta_i(t)=1$.
- (3) Assume $\alpha_i(t)=\alpha_{iq}$ and update $W_{iq}(t)$ and $Z_{iq}(t)$ according to the feedback signal $\beta_i(t)$.

$$\begin{cases} W_{iq}(t) = W_{iq}(t-1) + (1 - \beta_i(t)) \\ Z_{iq}(t) = Z_{iq}(t-1) + 1 \end{cases} \quad (3-8)$$

- (4) Update the optimum behavior of learning automata L_i according to the formula below and the optimum behavior refers to the one with the maximum value of $D_{ij}(t)$.

$$D_{ij}(t) = \frac{W_{ij}(t)}{Z_{ij}(t)} \quad (3-9)$$

(5) Assume the optimum behavior is α_{im} and update current learning automata L_i according to the behavior probability vector \mathbf{p}_i , where a denotes the award coefficient and p_j denotes the j^{th} component of vector \mathbf{p}_i .

$$p_j(t+1) = \begin{cases} p_j(t) + a(1 - p_j(t)) & j = m \\ (1 - a)p_j(t) & j \neq m \end{cases} \quad (3-10)$$

The procedure of the algorithm is described as below in detail:

- (1) Initialize the learning automata of each node randomly.
- (2) Each learning automata in the network select its own behavior according to self-behavior probability vector \mathbf{p}_i and obtain the community structure after decoding.
- (3) Every learning automata in the network learns and updates by interacting with local environment and overall environment.
- (4) Repeat step (2) until the community structure is stable.

Essentially, CLA-net algorithm is a multi-objective optimization algorithm with modularity function as its objective function and the condition that $k_i(c_i(t)) \geq k_i(c')$ where $c_i(t)$ denotes the community number which the node belongs to and $k_i(C) = \sum_{j \in C} A_{ij}$. A is the adjacent matrix of the network.

Apply the algorithm to the example in Figure 3-2 and detect community in the following procedure:

In the initial stage of algorithm, the state-transition matrix P can be calculated according to the adjacent matrix, where each element is the reciprocal of node degree.

$$P = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 & 0 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

The award matrix W and selection matrix Z are as follows.

$$W = \begin{pmatrix} 0 & 3 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 3 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 \end{pmatrix}, Z = \begin{pmatrix} 0 & 16 & 21 & 13 & 0 & 0 & 0 & 0 & 0 \\ 18 & 0 & 16 & 16 & 0 & 0 & 0 & 0 & 0 \\ 13 & 15 & 0 & 10 & 12 & 0 & 0 & 0 & 0 \\ 16 & 11 & 13 & 0 & 10 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 13 & 0 & 13 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 14 & 0 & 13 & 23 & 0 \\ 0 & 0 & 0 & 0 & 19 & 11 & 0 & 20 & 0 \\ 0 & 0 & 0 & 0 & 0 & 17 & 14 & 0 & 19 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 50 & 0 \end{pmatrix}$$

Calculate the value of matrix D according to matrix W and Z and let $D_i(t) = W_i(t)/Z_i(t)$. The optimum Q_{best} of community structure after the initialization is 0.3571.

The partition of community structure after an iteration is shown in Table 3-1.

Table 3-1 The partition of community structure after an iteration

Node number	1	2	3	4	5	6	7	8	9
Adjacent node	4	4	1	1	6	7	8	6	8

As shown above, node 1 and node 4 belong to the same community; node 2 and node 4 belong to the same community; node 3 and node 1 belong to the same community; node 4 and node 1 belong to the same community and so on. The partition of the community is shown in Table 3-2 after decoding.

Table 3-2 The partition of the community after decoding

Node number	1	2	3	4	5	6	7	8	9
Community number	1	1	1	1	2	2	2	2	2

The value of Q_{best} is 0.3571. All nodes are awarded because the value of Q_{best} remains the same and each node satisfies the condition $k_i(c_i(t)) \geq k_i(c')$. According to Formula (3-8), the award matrix W and selection matrix Z updated are as below.

$$W = \begin{pmatrix} 0 & 3 & 4 & 1 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 \end{pmatrix}, Z = \begin{pmatrix} 0 & 16 & 21 & 14 & 0 & 0 & 0 & 0 & 0 \\ 18 & 0 & 16 & 17 & 0 & 0 & 0 & 0 & 0 \\ 14 & 15 & 0 & 10 & 12 & 0 & 0 & 0 & 0 \\ 17 & 11 & 13 & 0 & 10 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 13 & 0 & 14 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 14 & 0 & 14 & 23 & 0 \\ 0 & 0 & 0 & 0 & 19 & 11 & 0 & 21 & 0 \\ 0 & 0 & 0 & 0 & 0 & 18 & 14 & 0 & 19 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 51 & 0 \end{pmatrix}$$

Then update matrix D according to the formula $D_i(t) = W_i(t)/Z_i(t)$. According to Formula (3-10), the updated state-transition matrix P is shown as below.

$$P = \begin{pmatrix} 0 & 0.2667 & 0.4666 & 0.2667 & 0 & 0 & 0 & 0 & 0 \\ 0.4666 & 0 & 0.2667 & 0.2667 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0.2 & 0 & 0.2 & 0.2 & 0 & 0 & 0 & 0 \\ 0.2 & 0.4 & 0.2 & 0 & 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0.2 & 0 & 0.4 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.4666 & 0 & 0.2667 & 0.2667 & 0 \\ 0 & 0 & 0 & 0 & 0.2667 & 0.4666 & 0 & 0.2667 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.4666 & 0.2667 & 0 & 0.2667 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

The results after several times of iteration is shown in Table 3-3.

Table 3-3 The community partition after several times of iteration

Node number	1	2	3	4	5	6	7	8	9
Community number	1	1	1	1	2	2	2	2	2

2. Multi-objective Optimization Algorithms on Overlapping Structure

The virtual community structure in online social networks is diverse, in which the overlapping community structure is a typical example. In order to detect virtual communities with overlapping structure in online networks, Jingfei Du et al. extended the existing multi-objective optimization algorithms to the field of overlapping community structure detection^[19]. In the algorithm, the topological structure of network is coded as gene sequence by edge-based mapping pattern to denote the situation that the same node belongs to several communities through edge clustering in network. The algorithm regard

the partition density function [Formula (3-11)] and the modularity function [Formula (3-11)] in overlapping communities proposed by Huawei Shen et al. as objective function which is defined as below.

$$D = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \quad (3-11)$$

$$Q_{OL} = \frac{1}{2m} \sum_{k=1}^c \sum_{i,j \in C_k} \frac{1}{O_i O_j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \quad (3-12)$$

The topological structure of network is coded as gene sequence based on adjacency of edge tracks. And the gene sequence can be optimized by genetic algorithms. Simulation results show that the algorithm can accurately detect the community structure in complex networks.

3.3.3 Algorithms Based on Probability Model

Bayesian probability model is widely used in topic detection initially^[24] and later applied in community detection algorithms. Probability model-based algorithms infer network model by maximizing Bayesian likelihood probability to obtain real network partition. The network model defines clustering structure by assuming the connection pattern among nodes and infer the best fit model to the observed network by maximizing Bayesian likelihood probability, thus obtaining the community structure. Some community detection algorithms based on probability model are introduced below.

1. Algorithms Based on Mixed Model

Mark Newman et al. propose a community detection algorithm in directed networks based on mixed model and expectation-maximization strategy. Given a directed network G with adjacency matrix A , its n nodes are partitioned into c communities. The mixed model parameters of such network is defined as follows.

g_i : the community of node i ;

π_r : the ratio between number of nodes in community r and number of nodes in network;

θ_{ri} : the probability that a directed edge from a certain node in community r to node i ;

According to definition, the sets $\{\pi_r\}$ and $\{\theta_{ri}\}$ satisfy the normalization condition:

$$\sum_{r=1}^c \pi_r = 1, \sum_{i=1}^n \theta_{ri} = 1 \quad (3-13)$$

The algorithm calculates the optimal value of parameters $\{\pi_r\}$ and $\{\theta_{ri}\}$ by fitting the mixed model to adjacency matrix A of observed network data.

According to defined parameters above, the connection probability between node i and j in mixed model is $p_{ij} = \pi_{g_i} \theta_{g_i, j}^{A_{ij}}$.

The likelihood probability that the observed network is generated by mixed model with parameters $\{\pi_r\}$ and $\{\theta_{ri}\}$ is

$$P(A, g | \pi, \theta) = P(A | g, \pi, \theta) P(g | \pi, \theta) \quad (3-14)$$

where

$$P(A | g, \pi, \theta) = \prod_{ij} \theta_{g_i, j}^{A_{ij}}, P(g | \pi, \theta) = \prod_i \pi_{g_i} \quad (3-15)$$

So the likelihood probability is

$$P(A, g | \pi, \theta) = \prod_i \pi_{g_i} \prod_{ij} \theta_{g_i, j}^{A_{ij}}$$

So the logarithm of likelihood probability $P(A, g | \pi, \theta)$ is

$$\mathcal{L} = \ln P(A, g | \pi, \theta) = \sum_i \left[\ln \pi_{g_i} + \sum_j A_{ij} \ln \theta_{g_i, j} \right] \quad (3-16)$$

The logarithm of likelihood probability is the objective function of the algorithm. The algorithm calculates the optimal value of parameters $\{\pi_r\}$ and $\{\theta_{ri}\}$ by maximizing the likelihood probability to determine the mixed model best fit to the observed network data. However, another unknown parameter g_i needs to be handled in the mixed model. For this purpose, a new parameter denoting the probability of node i in community r is defined as

$$q_{ir} = P(g_i = r | A, \pi, \theta) = \frac{\pi_r \prod_j \theta_{rj}^{A_{ij}}}{\sum_s \pi_s \prod_j \theta_{sj}^{A_{ij}}} \quad (3-17)$$

The expected value $\bar{\mathcal{L}}$ for the log-likelihood based on q_{ir} is calculated as

$$\bar{\mathcal{L}} = \sum_{ir} q_{ir} \left[\ln \pi_r + \sum_j A_{ij} \ln \theta_{rj} \right] \quad (3-18)$$

The parameters $\{\pi_r\}$ and $\{\theta_{ri}\}$ can be determined by maximizing expected value $\bar{\mathcal{L}}$ under the normalization condition of $\{\pi_r\}$ and $\{\theta_{ri}\}$, i.e.,

$$\pi_r = \frac{1}{n} \sum_i q_{ir}, \theta_{rj} = \frac{\sum_i A_{ij} q_{ir}}{\sum_i k_i q_{ir}} \quad (3-19)$$

where k_i is the out degree of node i . The final value of parameters $\{\pi_r\}$, $\{\theta_{ri}\}$ and $\{q_{ir}\}$ can be calculated by applying EM algorithm on Formula (3-17) and Formula (3-19). The steps of this algorithm are described below.

Input: the number of communities in network c , the maximal iteration number of algorithm $maxiter$.

(1) Initially, iteration number is set as $t=0$, and initial value of parameters $\{\pi_r\}_0$ and $\{\theta_{ri}\}_0$ are set as random values satisfying the normalization condition (3-13).

(2) Update values of parameter $\{q_{ir}\}_t$ by substituting current values of parameters $\{\pi_r\}_t$ and $\{\theta_{ri}\}_t$ into equation (3-17).

(3) Update values of parameters $\{\pi_r\}_{t+1}$ and $\{\theta_{ri}\}_{t+1}$ by substituting current values of parameter $\{q_{ir}\}_t$ into equation (3-19).

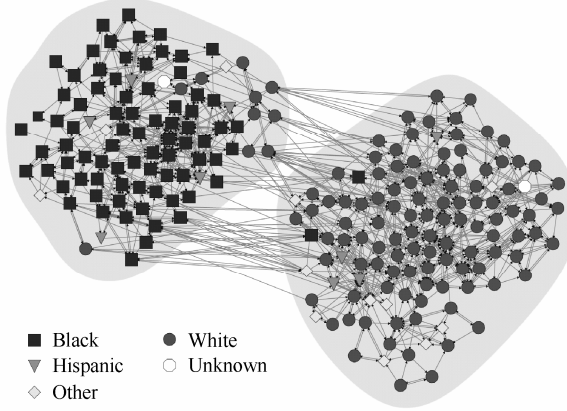
(4) Update iteration number as $t=t+1$. If $t > maxiter$, output the values of $\{q_{ir}\}_t$ and terminate the algorithm, otherwise go to step (5).

(5) Calculate the difference between values of parameter q in latest two iterations, i.e., $\Delta q = \|q_t - q_{t-1}\|$. If $\Delta q = 0$ which means the value of parameter q converges, end the algorithm and output the values of q , otherwise return to step (2) and continue iterations.

Note that the initial values of parameters π and θ should avoid unstable point of Formula (3-17) and Formula (3-19), i.e., $\pi_i = \frac{1}{c}$, $\theta_{ri} = \frac{1}{n}$.

The algorithm can be extended to deal with undirected networks naturally. In undirected network cases, the parameter θ_{ri} is defined as the connection probability between nodes in community r and node i . Other derivation processes are the same as those in directed network cases and finally obtain results of community portioning by Formula (3-17) and Formula (3-19). The outputting probability q_{ir} can be regarded as the belongingness intensity of node i in community r . One disadvantage of such algorithm is that the number of communities c needs to be set in advance, while such number is usually unknown.

Mark Newman et al. evaluate the algorithm in social network of U.S. high school students. The obtained community structure is illustrated in Figure 3-10. As shown in Figure 3-10, the algorithm partitioned the network into two communities, one of which mainly contains most black students in school and the other one mainly includes most white students. Students in other races uniformly belong to two communities, i.e. community structure of high school has strong relation to races of students.


 Figure 3-10 Social network of U.S. high school students^[22]

2. Edge Based Mixed Model Algorithm

Wei Ren et al. propose another mixed model based on edges to handle undirected networks. Given a undirected network G whose adjacency matrix is A , its n nodes are partitioned into c communities. The neighbor node set of a node i is denoted as $N(i)$. Similarly, the fraction of nodes in community r is π_r . Suppose that the probability that community r selects node i is β_{ri} which satisfies normalization condition $\sum_{r=1}^c \beta_{ri} = 1$. Larger value of β_{ri} indicates more importance of the node i in community r . Note that the node i can be selected by multiple communities. Assume that communities select different nodes independently, then the probability that the mixed model generates edge e_{ij} is

$$P(e_{ij} | \pi, \beta) = \sum_{r=1}^c \pi_r \beta_{ri} \beta_{rj} \quad (3-20)$$

The log-likelihood probability that such model generates all edges in network is

$$\mathcal{L} = \ln P(A | \pi, \beta) = \sum_{i=1}^n \sum_{j: j \in N(i)} \ln(\sum_{r=1}^c \pi_r \beta_{ri} \beta_{rj}) \quad (3-21)$$

Similarly, the log-likelihood probability can be maximized by EM algorithm. Belongingness intensity of each edge to different communities are obtained by maximizing log-likelihood probability and belongingness intensity of each node to different communities are derived accordingly.

3. Algorithm Based on LDA

LDA is a model for generating document topic and is also called three layer Bayesian probability model. LDA was first used to construct the three-layer model of word, topic and

document^[24]. Le Yu et al. extend LDA model to overlapping community detection tasks^[25] and propose an overlapping community detection algorithm LBLP based on LDA model. The algorithm includes three parts: network coding, edge LDA modeling, model inference. The time complexity of algorithm is relatively low, making it suitable for large online social network analysis.

3.3.4 Information Coding Algorithms

In order to compress the topology information, researchers introduced the idea of information coding in information theory and thereby designed new community detection algorithms. In information theory, information coding methods compress original information capacity by encoding more information with less codes according to the Minimum Description Length (MDL) principle. The main idea of MDL principle is that any rules in data can be used to compress data. The cohesive virtual community structure is an important law of data in online social networks. As a result, virtual community structure in online social network can be used to describe the information flow in network through compressive coding. Several algorithms are introduced as below.

1. Infomap Algorithm

Martin Rosvall et al. proposed the Infomap algorithm based on information theory^[5]. The algorithm uses random-walking as the agency of information spreading in network which will generate corresponding data flow. The amount of information generated by random-walking can be measured by the length of code in every step of random-walking, which is the average length of codes. Effective coding algorithms are needed for compressing the average length of codes up to the hilt. Huffman coding is a common coding algorithm which distributes short codes to each node visited by random-walking. Apply the Huffman coding to Figure 3-2 and the results as is shown in Figure 3-11. The information source coding theory of Shannon provides the theoretical boundary for the code length of Huffman coding: the average code length in every step should not be less than the entropy of variable X , i.e.

$H(x) = -\sum_1^n p_i \log(p_i)$. The sample space of variable X is set of nodes and the probability distribution of X is the visiting frequency of each node by

random-walking. As Huffman coding doesn't use the regularity of network structure, the average length of codes is still large. The second level coding underlining community structure can be considered to further compress the length of information flow.

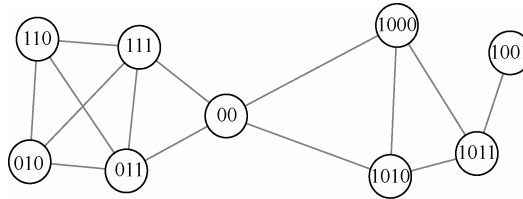


Figure 3-11 Coding on the example only by Huffman coding

Second level coding on the network in Figure 3-2, which is shown in Figure 3-12, allocates a unique codeword for each community in the network and different codewords for all the nodes in the same community. Codewords can be reused for nodes in different communities. This rule is similar to the naming rule on the map. Communities are similar to cities and nodes to streets in a city. Streets in different cities can share the same name while different streets in the same city cannot share the same name. Compared with the coding in Figure 3-11, this rule can be used to reduce the length of codeword effectively. The access frequency of communities or nodes can also be regarded as the probability distribution of variables, with codewords of communities and the nodes inside coded by Huffman coding. A leave code for previous community and a codeword for the community behind are needed in description in each random-walking between different communities to denote differences in communities. Second level coding algorithm converts the community partition problem to the problem of optimum coding, i.e. finding an optimum partition with minimal average description length in random-walking. The description length includes the codeword length in random-walking in communities and between communities. Obviously, better community partition leads to lower the shifting frequency between communities, reducing the average codeword length of communities. Moreover, the codeword length of node is greatly reduced because of the second level coding, leading to great compression of the overall length of description. On the contrary, the shifting will be more frequently and the length of random-walking description cannot be compressed if communities are not partitioned well.

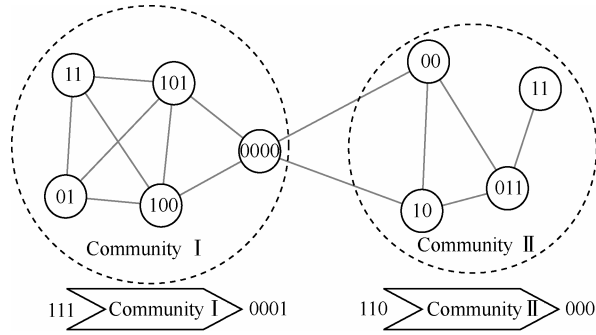


Figure 3-12 Second level coding on the example

Assume a community partition of given network with n nodes in the network partitioned into m communities. Then, the average description length $L(M)$ in every step of random-walking is expressed in the equation below.

$$L(M) = q_{\sim} H(Q) + \sum_{i=1}^m p_i^{\circ} H(p^i) \quad (3-22)$$

Formula (3-22) is called map equation, where $q_{\sim} H(Q)$ denotes the average codeword length of shifting between communities and $\sum_{i=1}^m p_i^{\circ} H(p^i)$ denotes the average codeword length in random-walking. The purpose of community detection in networks to find the optimal partition with minimal average codeword length.

Use an algorithm similar to the greedy algorithm in Section 3.3.1 to detect optimal partition, i.e. allocate a community for each node at first and combine two communities which lead to the most reduction in average description length $L(M)$. Repeat this process until communities are merged into one. The steps of the algorithm in detail is shown as below.

(1) Delete all edges in the network and regard each node as a community in the network.

(2) Consider each connected part as a community in the network and add edges outside the network back to the network. If the edge added connect two different communities, then combine these two communities. Calculate the decrement in average description length of the new partition and combine the two communities which lead to the most reduction in average description length.

(3) If the number of communities is more than one, then return to step (2) for iteration, otherwise go to step (4).

(4) Choose the community partition with minimal average description length as the optimal partition of network by traversing the values of average description length of

different community partitions.

Apply the Infomap algorithm to the example in this chapter with detailed steps shown as below.

(1) Regard each node as a community in the network, 9 communities in total, and calculate the average description length $L(M) = 5.132$.

(2) Combine any two communities and calculate the average description length of new communities. It is found that the combination of community $\{8\}$ and $\{9\}$ leads to the most reduction in average description length $\Delta L(M) = -0.3393$. As a result, the first step is to combine community $\{8\}$ and $\{9\}$ to community $\{8, 9\}$.

(3) Continue the process above until communities are merged into one. There are 9 different partitions in this process. The community spanning tree generated by the algorithm is shown in Figure 3-13.

(4) Traverse the values of average description length in different community partitions. It is found that the average description length has the minimal value $L(M) = 3.100$ if the network is partitioned into two communities $\{1, 2, 3, 4\}$ and $\{5, 6, 7, 8, 9\}$, which is the optimal partition of network.

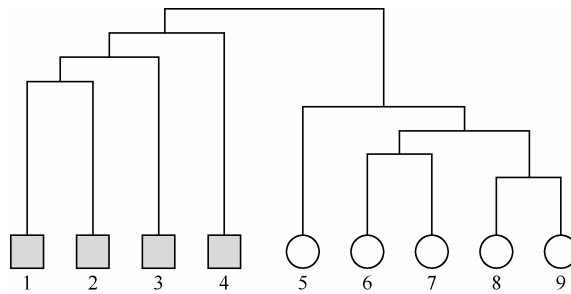


Figure 3-13 The community spanning tree generated by Infomap algorithm on the example

The Infomap algorithm detects the communities by the rule of information propagation in network. The algorithm is better than modularity optimization algorithm when the difference in size of communities is large. As the sizes of virtual communities in online social network is always different, Infomap algorithm is more accurate than modularity optimization algorithm.

2. The Map Equation of Edge Community

Original Infomap algorithm can only detect non-overlapping communities while online social networks usually have overlapping community structure. As a result, Youngdo

Kim et al. extended the map equation to detect edge communities in networks^[26]. Different from node communities, edge communities are partitioned by edges which partitions edges in networks into non-overlapping communities. The community of a node depends on its connected edge. As a node can be connected to several edges at the same time. A node will be partitioned into different communities if its connected edges are partitioned into different communities, which forms the overlapping community structure. Youngdo Kim et al. extended the map equation to detect edge communities in networks by modifying the coding rule of random-walking. The first level codes are allocated to edge communities in the network while the second level codes are allocated to nodes, and nodes belonging to different communities are allocated with different second level codes. As a result, the number of second level codes of each node is the same as the number of community which it belongs to. Random-walking will be still carried out on node and each step of random-walking starts from the source node to the target node going by an edge. If the edge in current step is different from that in previous step, record code of the current community and the secondary code of the target node. If these two edges are the same, record the second level code of the target node only. The map equation of edge community is shown as below.

$$L_{\text{linkcom}}(M) = q_{\sim} H(Q) + \sum_{i=1}^m p_{\circ}^i H(p^i) \quad (3-23)$$

Formula (3-23) and formula (3-22) are the same in expression, but are different in partition M and probability distribution q_{\sim} and p_{\circ}^i .

Although overlapping nodes have several second level codes which lead to redundancy, when the edge community partition is better than community partition of nodes, the utilization frequency of first level coding will be lowered along with the frequency reduction of random-walking between different communities. The reduction in average coding length can compensate for the redundancy of second level code. Besides, Youngdo Kim et al. thought that a network has important overlapping structure when the minimal description length L_{linkcom} in edge community partition is smaller than that in community partition of nodes. Therefore, they defined an index to measure the overlapping degree of communities in online social networks: the significance of overlapping, which is calculated as below.

$$\mathcal{O} = \frac{L_{\text{nodecom}} - L_{\text{linkcom}}}{L_{\text{nodecom}} + L_{\text{linkcom}}} \quad (3-24)$$

The value of the significance of overlapping satisfies the condition $\mathcal{O} \in (-1, 1)$. The overlapping structure is obvious if the value is positive.

Youngdo Kim et al. analyzed the blog network of U.S. politicians by map equation which has 1,490 nodes and 19,090 edges^[27]. By applying the map equation to node communities and edge communities respectively, they obtained the minimal description length in node communities of $L_{\text{nodecom}}=8.93$ and that in edge communities of $L_{\text{linkcom}}=8.65$. Moreover, the significance of overlapping of the network is $\mathcal{O}=0.0163$, which shows that online social networks, such as the Blog network, have obvious overlapping community structure.

3. Fast Algorithm by Minimizing the Map Equation

Fast algorithm to minimize the map equation is needed for online social networks of large scale. Martin Rosvall et al. proposed the fast algorithm by minimizing the map equation^[28] according to the fast modularity optimization algorithm^[17]. The core part of this algorithm is divided into two labels. In first stage, partition all nodes into independent communities and scan these communities randomly to combine the communities according to the largest reduction principle of map equation. Then start second stage after a round of scanning and reconstruct the network by regarding the communities obtained in first stage as new nodes in the network. Repeat these two labels alternatively until the value of map equation cannot be reduced any more.

Songchang Jin et al. proposed the community detection algorithm InfoMR^[29] based on MapReduce parallel framework by combining the information compression coding algorithm and parallel computation. The algorithm has low complexity and is suitable for online social networks of large scale because of the parallel computation mechanism.

3.4 Dynamic Calculation Detection Algorithms for Virtual Communities

3.4.1 Clique Percolation Algorithms

In online social networks, users can usually take part in different groups and topics, leading to an overlapping virtual community structure. The detection problem of overlapping communities is first proposed by Gergely Palla et al.^[6]. To solve this problem, some clique percolation algorithms are proposed according to the property that a node can belong to different cliques. Some clique percolation algorithms are

introduced as below.

1. CPM Algorithm

Cliques refer to complete subgraphs in networks which can also be called as groups. A subgraph with k nodes can be defined as k -clique^[6]. Internal edges of communities are easy to form clique due to high density of edges in the community and low density of edges between communities. Based on the characteristic above, Gergely Palla et al. proposed the CPM algorithm^[6]. Some basic concepts are introduced as below.

(1) Adjacent k -clique : two k -cliques are adjacent if they share $k-1$ nodes with each other.

(2) k -clique chain: a set of a series of continuous k -cliques is defined as a k -clique chain.

(3) The connectivity of k -clique : if two k -cliques are parts of the same k -clique chain, they can be deemed as connected.

(4) k -clique community: connected k -cliques in network, i.e. the set of all k -cliques which connect with each other by a series of adjacent k -cliques.

These concepts above can be explained with Figure 3-2 in this chapter. Assume $k=3$, 3-cliques in example include $\{1, 2, 3\}$, $\{1, 3, 4\}$, $\{3, 4, 5\}$, etc. Among all 3-cliques, $\{1, 2, 3\}$ and $\{1, 3, 4\}$ are adjacent because they share node 1 and node 3. $\{1, 2, 3\}$, $\{1, 3, 4\}$ and $\{3, 4, 5\}$ form a 3-clique chain in which $\{1, 2, 3\}$ and $\{3, 4, 5\}$ are 3-clique connected. At last, the set $\{1, 2, 3, 4, 5\}$ forms a 3-clique connected part called as a 3-clique community. According to the steps above, another 3-clique community in example is $\{5, 6, 7, 8\}$. As shown in figure, the communities partitioned are overlapped as these two 3-clique communities include node 5 at the same time.. If k equals 4 in this example, the 4-clique community is $\{1, 2, 3, 4\}$. Obviously, if the value of k is larger, the size of k -clique community will be smaller but the communities will be more cohesive.

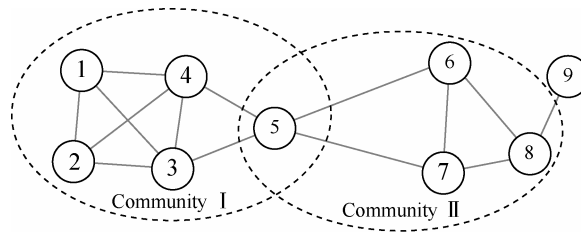


Figure 3-14 The overlapping community structure of example detected by CPM algorithm

Parameter k is a threshold value for complete subgraphs. In fact, complete subgraph with size s larger than k may exist. Obviously, a complete subgraph with size s will include C_s^k different k -cliques and they are k -cliques connected. If two large complete subgraphs share at least $k-1$ nodes, they are also k -clique connected and form a k -clique community. In the example, set $\{1, 2, 3, 4\}$ is a 4-clique which includes four 3-cliques which are 3-clique connected. This 4-clique share two nodes with the 3-clique $\{3, 4, 5\}$. So they are 3-clique connected and form a 3-clique community. To describe the algorithm simply, the cliques not belonging to other complete subgraphs will be called as the maximum clique. CPM algorithm detects overlapping communities by seeking the maximum cliques in networks of which the size is not smaller than k .

The steps of the algorithm is described as below.

Input: The threshold value for the size of clique is k .

(1) Find all the maximum cliques in the network.

(2) Construct the clique-clique overlapping matrix O . O is a symmetric matrix in $n_c \times n_c$ order. n_c denotes the number of maximum clique in the network. O_{ij} denotes the number of nodes shared by maximum clique i and j . O_{ii} denotes the size of the maximum clique.

(3) Non-diagonal elements smaller than $k-1$ and diagonal elements smaller than k are all set to 0, and the remaining elements are set to 1.

(4) Analyze the processed matrix O and find the connected part which is regarded as the final k -clique community.

Apply the algorithm to the example.

Input: The threshold value $k=3$.

(1) The maximum cliques in the network are $\{1, 2, 3, 4\}$, $\{3, 4, 5\}$, $\{5, 6, 7\}$, $\{6, 7, 8\}$, $\{8, 9\}$.

(2) Construct the clique-clique overlapping matrix of the network.

$$O = \begin{bmatrix} 4 & 2 & 0 & 0 & 0 \\ 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 3 & 2 & 0 \\ 0 & 0 & 2 & 3 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix}$$

(3) Non-diagonal Elements smaller than 2 and diagonal elements smaller than 3 are set to 0, and the remaining elements are set to 1.

$$O = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(4) Analyze the processed matrix and find two connected part $\{1, 2, 3, 4, 5\}$ and $\{5, 6, 7, 8\}$. These parts are two 3-cliques in the network with the overlapping node 5.

Detecting the maximum clique in the network is a NP complete problem. Gergely Palla et al. put forward an effective algorithm to detect the maximum clique in networks. Firstly, specify the size s of the maximum clique according to the degree sequence of network nodes; secondly, select nodes repeatedly and extracts all the maximum cliques containing the node with size s ; thirdly, delete the nodes selected and their connected edges until there are no nodes left; at last, reduce the size by 1 and detect new maximum cliques in the network. Repeat this process above until the size of maximum clique is 0. The algorithm is fast in real networks because real networks are sparse and have limited maximum cliques.

In online social networks, users tend to form cliques. Large communities usually have many small groups in which users are related or close to each other and form the cliques in communities. At the same time, each user belongs to several groups and forms the overlapping structure in communities. As a result, detection algorithms on overlapping community structure based on cliques are suitable for community detection in online social networks.

2. CPMw Algorithm

In terms of weighted online social networks, a standard algorithm is to set a threshold value for the edge weight. In this case, edges with weight less than the threshold value are excluded from the network and the remaining are regarded as unweighted edges which form a corresponding unweighted network. Finally, apply the CPMw algorithm on this unweighted network. Illés Farkas et al. proposed another weighted clique percolation algorithm^[30]. They defined the intensity of k -clique according to the intensity of subgraph, where the intensity of a subgraph refers to the geometric average of all the

weights of edges in the subgraph^[31]. As a result, the intensity of k -clique \mathcal{C} can be defined as below.

$$I(\mathcal{C}) = \left(\prod_{i < j, i, j \in \mathcal{C}} w_{ij} \right)^{2/k(k-1)} \quad (3-25)$$

Different from basic algorithm, the algorithm sets a threshold for the intensity I . k -clique with intensity larger than the threshold will be partitioned into the community; otherwise, it will be abandoned. Other steps in the algorithm is similar to CPM algorithm. The algorithm allows for edges with weight smaller than I in k -cliques and considers the integrality of the topological structure in weighted online social networks, which avoids the situation that some users are partitioned into different communities by mistake because of their weak connection.

3. Fast Clique Percolation Algorithms

Jussi Kumpula et al. proposed a fast clique percolation algorithm named SCP^[32] for community detection in online social networks of large scale. The algorithm includes two stages. In the first stage, it aims to find k -cliques in networks. Starting from an empty graph, the algorithm adds edges into the network one by one and inspects that if a new k -clique is generated. In second stage, regard the k -cliques generated above as input and judge whether k -cliques generated in first stage belongs to existing k -cliques by calculating the overlapping degree of them. A final partition will be generated by these two steps.

Time complexity of the algorithm is linear to the number of k -cliques and the run time is much shorter than that of original CPM algorithm especially in weighted networks. Original CPM algorithm needs to operate the algorithm for all the threshold values of weight, while SCP algorithm is operated only once, saving lots of time. In terms of weighted social networks of large scale, the best threshold for the weight of relationship between users are uncertain, SCP algorithm can obtain overlapping community structure of all threshold values so as to find the best threshold value and best community structure.

3.4.2 Agglomerative Algorithms Based on Similarity

In real world, the community structure is usually organized in a hierarchical form. Take colleges as example, faculty and students can be partitioned into several

communities according to their institutes. All the members can also be partitioned into different departments while students in different classes of each department form community structure too. As abstraction and extension of real social relationship, online social networks are also organized in the hierarchical form. For example, on Facebook and other social networks aboard, staff from the same company can form community structure according to their departments. While each community contains several sub communities. Online social networks in China such as renren.com have similar hierarchical feature in community structure which has great importance in understanding of the topological feature of the entire network and mining the function of each module in networks. For this purpose, many scholars design community detection algorithms to reveal the hierarchical community structure in topological structure of networks. Among these algorithms, agglomerative algorithms based on similarity are of great importance. They referred the idea of clustering in traditional pattern recognition. In initial stage, regard each node as an independent community; then merge two sub communities with biggest similarity into a bigger community until all nodes are merged into the biggest virtual community.

1. EAGLE Algorithm

Virtual community aggregation algorithm based on similarity can easily detect a virtual community with a hierarchical structure and reveals structural characteristics of online social networks organized at different levels. Huawei Shen et al. proposed a typical hierarchical community aggregation algorithm - EAGLE algorithm^[14], which detects all of the maximum clique with the technology of detecting the maximum clique. On this basis, use the traditional data clustering framework and iteratively merge sub-communities with the greatest similarity until the entire network is merged into a big community, where the degree of similarity between the sub-communities is expressed by the following formula:

$$M = \frac{1}{2m} \sum_{v \in C_1, w \in C_2, v \neq w} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \quad (3-26)$$

C_1 and C_2 denotes two communities, A_{vw} denotes the element of the network adjacency matrix, m denotes the number of edge in network diagram A , and k_v , k_w denotes the degree of node v and node w respectively.

Specific process is as follows:

- (1) According to the network structure of the current input, detect the largest clique in

the network with the Bron-Kerbosch algorithm; then filter the smallest scale clique according to a preset threshold k , and consider the maximum clique of the community with scale larger than k as initialized community structures, and calculate the similarity between communities using the Formula (3-26).

(2) Select two community structures which have maximum similarity and merge them into one community; then calculate the similarity of the new community and other communities.

(3) Repeat step (2) until the entire network merge into one large community structures.

(4) Determine the largest structure of the community partition which can make Formula (3-27) obtain the maximal value and output it as a result. Formula (3-27) is as follows:

$$EQ = \frac{1}{2m} \sum_i \sum_{v \in C_i, w \in C_i} \frac{1}{O_v O_w} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \quad (3-27)$$

O_v is the number of the community which node v belongs to. Community partition obtaining the largest EQ is the optimal partition of community structure.

The time complexity of EAGLE algorithm is as follows: For a network with n nodes, assume that the maximum number of clique is s in the initial state, h is the number of the maximum cliques sharing an edge, then step of the algorithm time complexity is $O(n^2)$ in step (1); in step (2), the merge of largest clique need to be operated $s-1$ times. Every merge operation needs to partition $h+n$ communities and calculate similarity between new communities and other communities, so the time complexity is $O(h+n)$, therefore the first three steps of total time complexity is $O(n^2 + (h+n)s)$. According to Formula (3-27), the actual complexity of the step (4) is $O(n^2 s)$. Therefore, the real complexity is $O(n^2 + (h+n)s + n^2 s)$.

For example, detect the maximum clique in the network diagram by EAGLE algorithm in Figure 3-2. There are four maximum cliques in the graph: $\{1, 2, 3, 4\}$, $\{3, 4, 5\}$, $\{5, 6, 7\}$ and $\{6, 7, 8\}$. Based on the EAGLE algorithm, the isolated points except the maximum clique are also considered as a separate clique in the network, such as $\{9\}$. By Formula (3-26), we can calculates that the similarity between $\{5, 6, 7\}$ and $\{6, 7, 8\}$ as 0.0995. For all the neighboring maximum clique, it has the greatest similarity. Therefore, choose the biggest clique consolidation, merging $\{5, 6, 7\}$ and $\{6, 7, 8\}$ into the same community. After the second iteration, the similarity between the cliques $\{1, 2, 3, 4\}$ and $\{3, 4, 5\}$ is 0.0714, while the similarity between cliques $\{3, 4, 5\}$ and $\{5, 6, 7, 8\}$ is -0.0561. Therefore,

in the second iteration, merge $\{1, 2, 3, 4\}$ and $\{3, 4, 5\}$ into $\{1, 2, 3, 4, 5\}$. In the third iteration, the similarity between cliques $\{5, 6, 7, 8\}$ and $\{1, 2, 3, 4, 5\}$ is -0.1556 , while similarity between cliques $\{5, 6, 7, 8\}$ and $\{9\}$ is 0.0191 , so merge clique $\{5, 6, 7, 8\}$ and $\{9\}$. Finally, $\{1, 2, 3, 4, 5\}$ and $\{5, 6, 7, 8, 9\}$ are merged into one large clique. Clique merging tree is as shown in Figure 3-15. According to Formula (3-27), it is most reasonable when the community is partitioned in the position shown in dotted lines.

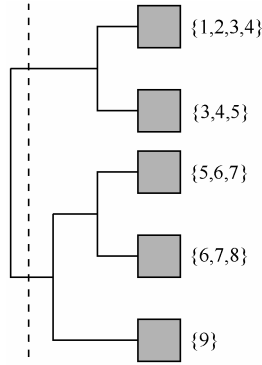


Figure 3-15 Clique merging tree in EAGLE algorithm

2. Other Agglomerative Algorithms Based on Similarity

The most important part of aggregation algorithm based on the similarity is to calculate the similarity between the sub-communities. Using different similarity function has a certain influence on the final result of community partition. Over the past few decades, scholars in various fields proposed various similarity formulas for this problem and depicted and measured the degree of similarity between different nodes from different perspectives. These similarity formulas mostly map sub-community structures to nodes in n -dimensional space, then measure the degree of similarity between these sub-communities by distance between nodes and other concepts. A typical concepts is the Euclidean distance, M distance, infinite norm and cosine similarity formula.

For two nodes $A=(a_1, a_2, \dots, a_n)$ and $B=(b_1, b_2, \dots, b_n)$, formulas based on the similarity of Euclidean distance, M distance and infinite norm are as below.

$$d_{AB}^E = \sum_{k=1}^n \sqrt{(a_k - b_k)^2} \quad (3-28)$$

$$d_{AB}^M = \sum_{k=1}^n |a_k - b_k| \quad (3-29)$$

$$d_{AB}^\infty = \max_{k \in [1, n]} |a_k - b_k| \quad (3-30)$$

Experts and scholars of virtual community use the knowledge of geometry and statistics, then measure the similarity degree between communities based on cosine distance as below

$$\rho_{AB} = \arccos \frac{\mathbf{A} \cdot \mathbf{B}}{\sqrt{\sum_{k=1}^n a_k^2} \sqrt{\sum_{k=1}^n b_k^2}} \quad (3-31)$$

where $\mathbf{A} \cdot \mathbf{B}$ denotes the inner product of vectors \mathbf{A} and \mathbf{B} .

In addition to using space node information, we can also use the network topological structure information to calculate similarity between sub-communities. As community structure is a set of nodes connected densely inside, many scholars measure the degree of similarity between the sub-community networks via neighbor nodes of information characteristic of this topology, where a typical example is the similarity based on the number of direct neighbor nodes expressed as below.

$$\omega_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \quad (3-32)$$

$\Gamma(i)$ and $\Gamma(j)$ denote collection of a direct neighbor node of i and j respectively. $|X|$ denotes the number of elements in the set X .

The above formula depicts the similarity degree between the community structure from different perspective and sides. Selecting the appropriate similarity formula according to different network characteristics can effectively improve the detection accuracy of the community. For example, Jianbin Huang^[40] et al. consider each network node as community structure which has a node and use the similarity formula [Formula (3-31)] based on cosine representing the similarity between communities of the similarity between communities and detect connected closely nodes in the local area by the local notion of the strongest edges for merging. Each combined community will be deemed as a virtual node and involved to subsequent merger process.

3.4.3 Label Propagation Algorithms

Label propagation algorithms, proposed by Xiaojin Zhu et al. in 2002, is a

semi-supervised learning algorithm based on graph which forecasts the information of unlabeled nodes according to information of labeled nodes^[33]. Since the algorithms are simple and have low time complexity, Usha Nandini Raghavan et al. applied them to detect communities in networks^[8]. These algorithms need no specific objective functions and define community structure by intuitive heuristic rules. Some label propagation algorithms are introduced as below.

1. LPA Algorithm

The basic idea of LPA algorithm is to set labels for all nodes in networks and design a propagation rule through which labels are propagated iteratively until all labels are stable. Then nodes with the same label are partitioned into the same community. The label of each node is updated as the label with most neighbor nodes. The propagation rule defines the community structure of a network, in which each node is partitioned into the community that most neighbor nodes belong to.

Steps of the algorithm is shown as below in detail.

- (1) At the beginning, initialize each node by a unique label.
- (2) The label of each node is updated as the label with most neighbor nodes by scanning all nodes in a random order. If there are several labels with most neighbor nodes, select one randomly.
- (3) If the label of each node is the same as the label with most neighbor nodes, go to step (4); otherwise, return to step (2).
- (4) Finally, regard each connected part with the same label as a community.

Apply the LPA algorithm to the example in this chapter. The propagation process of labels is shown in Figure 3-16, which can be described as below.

- (1) In the beginning, label each node by its own node number.
- (2) Generate a random order {1, 8, 6, 5, 7, 3, 2, 4, 9} for all nodes. Then scan all nodes in this order and update labels according to the updating rule.
- (3) After scanning in this order, the label of each node is the same as the label with most neighbor nodes, i.e. all labels in the network are stable.
- (4) According to different labels of nodes, the network can be partitioned into two communities, i.e. {1, 2, 3, 4} and {5, 6, 7, 8, 9}.

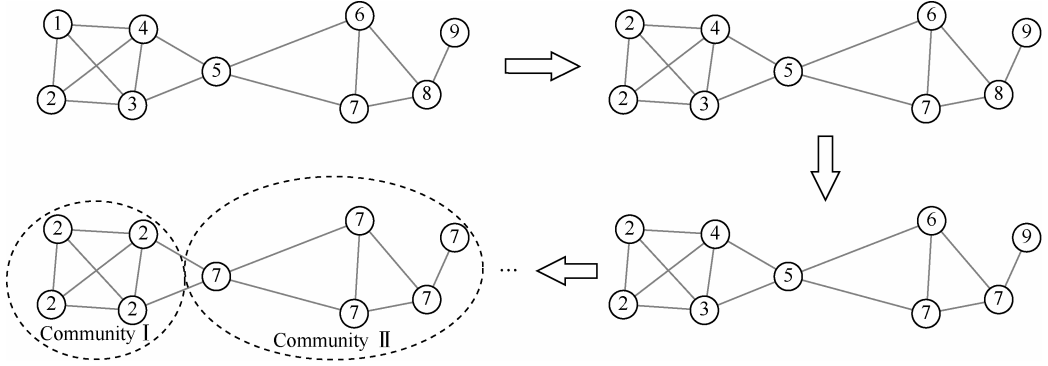


Figure 3-16 The propagation of labels on the example

In this algorithm, to avoid the algorithm cycle and ensure the convergence, reorder nodes randomly and update labels asynchronously before each propagation. The community structure may be non-unique according to this rule. Several community structures may satisfy the stop condition according to the same initial condition. But they are similar to each other. Figure 3-17 shows two possible community structures of the example. Node 5 has two neighbors in both two communities. As a result, it can be partitioned into either community I or community II. A community structure with more information can be generated by merging labels of a node in different community structures into one.

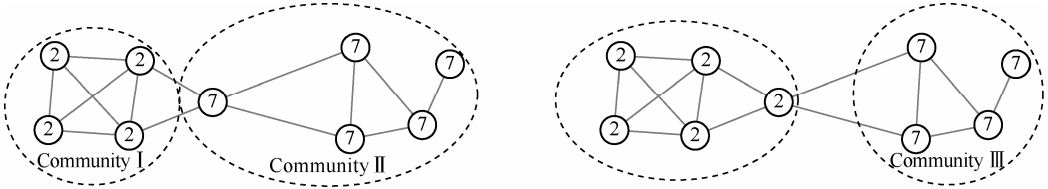


Figure 3-17 Two possible community structures of the example

The algorithm can detect communities naturally according to the topological structure of networks without any parameters including the number and the size of communities. Besides, the time complexity is almost linear. The time complexity for initializing each node is $O(n)$. In each iteration, for each node x , group its neighbors by their labels; then update the label of node according to the label of its largest neighbor group. This process costs $O(d_x)$, where d_x denotes the degree of node x . Repeat this process for each node and the time complexity in each iteration is $O(n\bar{d})$, i.e. $O(m)$. Experiments show that the number of iteration need for the convergence of algorithm is usually independent from the

size of a network. In general, 95% nodes will be partitioned correctly after 5 iterations. Therefore, the time complexity of this algorithm is very low which is approximately $O(n)$ on sparse networks. The core idea of this algorithm is similar to the formation of communities in online social network to some extent. In online social networks, users tend to participate in the same subject as its most neighbors, which forms communities in networks. Appearance of communities only depends on the local information of networks. As a result, LPA algorithm is suitable for community detection in large online social networks.

2. An Extension of LPA Algorithm for Overlapping Communities

In original LPA algorithm, a node has only one label in the process of label propagation. So a node belongs to only one community finally and the community structure discovered by the algorithm is non-overlapping. In order to fit for the overlapping community structure in online social networks, Steve Gregory extended LPA algorithm and proposed the COPRA algorithm^[34].

The algorithm improved the process of label propagation. It allows for multiple labels for each node to contain the information of multiple communities. To give each node accurately with multiple labels, Steve Gregory provided a belongingness coefficient b of each label c for each node, which composes a relationship pair (c, b) . Belongingness coefficient b also denotes the intensity of belongingness of a node to community c . As a result, all coefficients need to be normalized. In the process of label propagation, labels of each node are updated as the union sets of its neighbor labels, then normalize the belongingness coefficients of each node. As shown in Figure 3-18, some of the nodes $\{3, 4, 5, 6, 7\}$ in the network are selected to operate the label propagation. In the beginning, each node has only one label with belongingness coefficient 1. Labels of each node will be updated as the union sets of its neighbor labels and the belongingness coefficients of each node will be normalized after an iteration. In this case, each node can keep all the labels. In order to remove those unimportant labels, a threshold value ν can be set to delete the labels with belongingness coefficient smaller than $1/\nu$ to ensure each node has ν labels at most, i.e. it can belong to ν communities at most. As a result, ν also denotes the largest number of communities that a node belongs to.

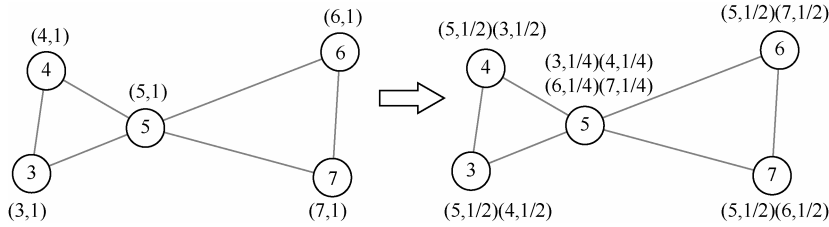


Figure 3-18 A propagation iteration of labels in the network composed of $\{3, 4, 5, 6, 7\}$

Time complexity of the algorithm in each iteration on sparse networks is $O(v^3n + vn\log(v))$. For an online social network, v denotes the largest number of communities that a node belongs to, which is a constant much smaller than the size of networks n . Therefore, the time complexity of the algorithm in each iteration is almost linear.

3. Improving the Stability of LPA Algorithm

Although LPA algorithm is simple and has low time complexity, it is an uncertainty algorithm. Ian Leung et al. found that, after using asynchronous updated LPA algorithm for many times, a variety of community structure may be obtained for the same online social network. In some structures, there is a community with particularly large scale. The community may contain more than 50% of the nodes in a network^[35]. Because when the edge density in some communities is not high enough, its internal label may have been invaded by other communities, which generates a huge community.

To solve this problem, Ian Leung et al. put forward an algorithm to improve the LPA algorithm^[35]. The original updating rule of LPA does not consider the propagation distance. No matter how far the label propagate, the effect on the updating of other labels is invariable. This may lead to the result that the label in a community can propagate very far so as to invade other communities and generate a huge community. In the improved algorithm, Ian Leung et al. assigned a score for each label which decreases with the increase of label propagation distance. Using the score to weight the influence of label's propagation process, the influence of scores to the update other labels will gradually become smaller along with the increase of label propagation distance, thus effectively preventing a label from propagating too far and invading other communities.

Another important reason of the instability of LPA algorithm is the randomness in the process of the algorithm. For example, when label is updated asynchronously, each iteration of algorithm needs a new random sequence. Therefore, when executing the algorithm on

the same data set for multiple times, the update order of nodes is different, which may lead to multiple different community structures. Yuxin Zhao et al. designed a label propagation algorithm LPA-E based on the entropy order. The algorithm uses the label entropy in an ascending order in each iteration of label propagation and removes the randomness in original algorithm so as to make the community structure more stable^[36].

In order to make the community structure more stable and reasonable, Hao Lou et al. put forward an improved LPA algorithm based on Coherent Neighborhood Propinquity^[37]. When updating node labels, they introduced the CNP between nodes to measure the propinquity of any node pair in networks. When updating the label, they used CNP between nodes to weight node labels. Two nodes are more close to each other and the mutual effect between their labels is greater as the CNP value between them is larger. When updating labels, the label with biggest CNP value is updated. The idea that nodes who are more close to each other have greater mutual influence in the process of label propagation also conforms to online social networks. Users who are more close to a certain user have greater influence on community selection of the user, i.e. users tend to choose the community that their intimate partners participate in.

3.4.4 Local Expansion Optimization Algorithms

Virtual community structure is local structure in online social networks. The formation of a virtual community only depends on connecting relation of local network and has nothing to do with topological structure in other areas of the network. Therefore, the algorithm based on local topological information is more in line with the characteristics of the virtual community in online social networks. This algorithm defines a health function based on the local topological structure of network. It starts with a seed community and iteratively extends the seed community until the health function is optimized which forms an optimal natural community. Natural communities extended by different seed communities overlap with each other. Therefore, this algorithm can detect the overlapping community structure of networks and is suitable for detecting overlapping communities in online social networks. Several local expansion optimization algorithms are described as below.

1. LFM Algorithm

Andrea Lancichinetti et al. put forward LFM algorithm according to the optimization of the local expansion. The algorithm starts from a seed community and iteratively

discovers natural communities of all nodes which forms the final partition of communities in networks. Natural community of nodes is defined as the subgraph which has the largest health degree. That is to say adding a new node in the subgraph or deleting a node from the subgraph will decrease the health degree of the subgraph. The health degree of subgraph \mathcal{G} is defined as below.

$$f_{\mathcal{G}} = \frac{\kappa_{\text{in}}^{\mathcal{G}}}{(\kappa_{\text{in}}^{\mathcal{G}} + \kappa_{\text{out}}^{\mathcal{G}})^{\alpha}} \quad (3-33)$$

In this formula, $\kappa_{\text{in}}^{\mathcal{G}}$ and $\kappa_{\text{out}}^{\mathcal{G}}$ respectively denote the sum of internal and external degree of all nodes in subgraph \mathcal{G} . α is a positive real parameter used to control the size of discovered community. According to health degree of the subgraph, the health degree of node A in a subgraph \mathcal{G} is defined as below.

$$f_{\mathcal{G}}^A = f_{\mathcal{G}+\{A\}} - f_{\mathcal{G}-\{A\}} \quad (3-34)$$

In this formula, $f_{\mathcal{G}+\{A\}}$ and $f_{\mathcal{G}-\{A\}}$ respectively denote the health degree including node A and without node A in subgraph \mathcal{G} .

Specific steps for discovering natural community of node A are as follow.

Input: the value of parameter α .

- (1) Initially, natural community \mathcal{G} of node A only includes A ;
- (2) Implement a cycle to all neighbors which are not in \mathcal{G} and calculate the health degree of each node with respect to \mathcal{G} . If the maximal health degree is positive, add the node which has the maximum health degree into \mathcal{G} . Otherwise, the algorithm ends and output the natural community \mathcal{G} of A ;
- (3) Recalculate the health degree of all nodes in \mathcal{G} ;
- (4) If the minimal health degree is negative, delete the node with minimal health degree from \mathcal{G} and turn back to step (3). Otherwise turn back to step (2).

Steps of finding natural community of node 1 in the example in this chapter by the algorithm are as follow.

Input: parameter $\alpha = 1$.

- (1) Initially, natural community \mathcal{G} of node 1 only includes node 1;
- (2) Implement a cycle to all neighbor nodes $\{2, 3, 4\}$ which are not in \mathcal{G} and calculate all neighbor nodes' health degree with respect to \mathcal{G} . It is found that the maximal health degree is 2 and $f_{\mathcal{G}}^2 = 0.33$. Thus, add node 2 into \mathcal{G} .
- (3) Recalculate the health degree of each node in \mathcal{G} ;
- (4) At this time, the health degree of each node in \mathcal{G} is positive. Then, continue to

calculate the health degree of each neighbor node with respect to \mathcal{G} .

Expansion process of natural community of node 1 is shown in figure 3-19. T denotes the number of iterations in expansion process. When \mathcal{G} is $\{1, 2, 3, 4, 5\}$, the maximal local health degree is $f_{\mathcal{G}} = 0.89$.

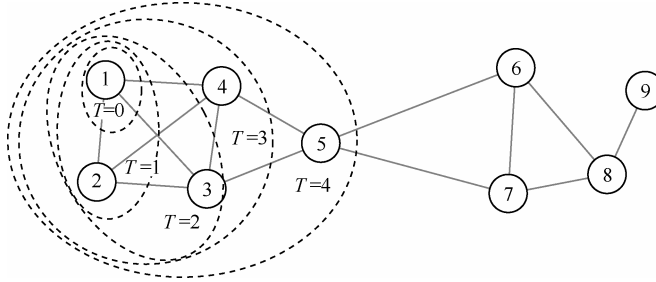


Figure 3-19 Expansion process of natural community of node 1

The parameter α in the health degree function is very important for it can control the resolution to find communities. Large α will result in small communities, and for small α the situation is reversed. Existed experiments shows that, under most circumstances, if $\alpha < 0.5$, the network can only be partitioned into a single community. However, if $\alpha > 2$, we can get minimal community structure in networks. A natural way of choosing value for α is 1, which can make the definition of health degree and weak community structures the same^[38]. Under most circumstances, the corresponding community structure with $\alpha = 1$ is relatively more important for it reveals key information of real network community structures. Due to the fact that different parameter resolution correspond to different community scales, this algorithm can be adopted to explore community structures in different levels.

Similar to label propagation algorithm, this algorithm is a non-deterministic one as well. Because it may find different community structures when selecting seed node with different orders. The execution time of it largely depends on the community scale and degree of overlapping between communities. To establish a natural community with scale of s , the algorithm requires a time complexity of approximately $O(s^2)$, therefore for a fixed value α , the time complexity of the algorithm is $O(n_c \langle s^2 \rangle)$, where n_c is the number of natural communities and $\langle s^2 \rangle$ is the second order moment. So the algorithm has approximately linear time complexity as the scale of community is relatively smaller than scale of networks. With respect to online social communities, the scale of virtual community is generally smaller than the scale of networks, thus resulting in a relatively small time complexity of the algorithm, so this algorithm is suitable for analysis in large scale networks.

2. GCE Algorithm

LFM algorithm utilizes nodes as seeds to expand communities. It deletes nodes with negative health degree during the expanding procedures. To reduce the computational complexity, LFM only find the natural communities of non-distributed nodes, instead of finding communities of all nodes. However, in this way, the communities of each node may be few, which leads to the low degree of overlapping in the community structures. In fact, overlapping degree of communities depends on the number of nodes that belong to different communities simultaneously. In social networks, every user can join in arbitrary amount of virtual communities based on their personal interests. Thus there exists high overlapping degree in online social networks. However, most overlapping community detection algorithms will be effective only under the circumstance that the degree of overlapping is low. With respect to this problem, Conrad Lee et al. suggested GCE algorithm (another local expansion optimization algorithm), which is more suited for highly overlapped communities^[39].

GCE algorithm find a set of seeds at first and then expand these seeds to form a community by using greedy algorithm to optimize a local health function. At last, it accepts those communities that are not similar to existed ones. GCE algorithm applies the same local health function as the one used by LFM algorithm. In contrary to LFM algorithm, this algorithm use maximum cliques as seeds. In the content that follows we will call maximum clique as clique for short. Inside the implementation procedure of the algorithm, small cliques is not desired as seeds, so a threshold value k should be used to discard unqualified cliques. After getting all cliques, the algorithm put all neighbor nodes into cliques with the help of greedy local optimization health function until corresponding health function of subgraph reaches maximum. By expanding all groups to form communities, this algorithm can find highly overlapped community structures. But different groups may result in same or similar communities. To dealt with the same or similar communities, this algorithm discard new communities that are same as or similar to existed communities during the procedure when finding seed communities sequentially.

3.5 Summary

Recently, with the wide development of social network services like Facebook, Twitter, Sina Weibo, etc., online social networks gradually become the focus of scholars from various fields. Virtual community detection in online social network is an important part in

social networks analysis. Analyzing the community structure and composition in social networks helps in researching characteristics of topological structure of social network, finding user's clustering patterns and influence factors, boosting information indexing, information recommendation, information propagation & control, public security events and other applications. Therefore, research of community detection algorithms has important social meaning and application value.

Based on various characteristics of online social networks and the traditional complex network community theory, this chapter summaries the definition of community structure and partitions these mainstream community detection algorithms into static and dynamic calculation detection algorithms. Finally, we introduced related research works in virtual community detection algorithms.

Even if algorithms and technologies of community detection have reached tremendous theoretical and application success, we think it still requires deeper research in following aspects:

(1) At present, majority of community detection algorithms are proposed against static social network, i.e. the network structure doesn't change along with the lapse of time. However, as the matter of fact, online social network is a continuous changing dynamic network. New nodes and edges are added continuously. Thus, a more challenging task is to find community structure in dynamic social networks. An intuitive approach in dynamic social networks is to enforce snapshot, which is mapped to certain static topological structure at a certain time point. This approach can find community structures in social networks at different time and detect the evolution law of structure in networks. However, it has bad resistance against noise in networks, and the algorithm has high time complexity for it needs to be operated on each snap. As a result, it is not suitable for community detection in online social networks in time. Therefore, it's necessary to develop efficient virtual community detection approaches and technologies in dynamic social networks based on 'increment' data in networks.

(2) Most virtual community detection algorithms available currently are based on modifications of traditional algorithms in complex networks, which is only applicable to homogeneous network structures, i.e., networks with only a single type of nodes and edges. As products of the combination of real social networks and internet technology, edges and nodes of online social networks may have multiple types. For example, nodes can denote users, information (like photos, labels, etc.), and edges can also denote the relationships between users, the relationships between users and labels, the relationships between photos

and labels, etc. Different kinds of nodes and edges provide us with rich and valuable information. However, it also brings us the obstacle in how to deal with them properly. As a result, the diversity of nodes and edges brings us new challenges as well as opportunities in developing excellent virtual community detection algorithms.

References

- [1] Mark Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 066133, Jun. 2004.
- [2] Brian Wilson Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs [J]. *Bell.Syst.Tech.j.*, 1970, 49: 291-307.
- [3] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory [J]. *Czech Math.J.*, 1977, 25: 619-633.
- [4] Mark Newman and Michelle Girvan. Finding and evaluating community structure in networks [J]. *Physical Review E*, 69, 026113, Feb. 2004.
- [5] Martin Rosvall and Carl Bergstrom. Maps of random walks on complex networks reveal community structure [J]. *Proc. Natl. Acad. Sci.*, 2008, 105(4) : 1118-1123.
- [6] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society [J]. *Nature*, 2005, 435 : 814-818.
- [7] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure of complex networks [J]. *New J. Phys.* 11, 033015, Mar. 2009.
- [8] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks [J]. *Physical Review E*, 76, 036106, Sep. 2007.
- [9] Santo Fortunato. Community detection in graphs [J]. *Physics Reports*, 486, Feb, 2010.
- [10] Leon Vicsek Danon, Jordi Duch, Alex Arenas, Albert Diaz-Guilera. Comparing community structure identification [J]. *J.Stat. Mech.Theory Exp.*, 09008, 2005.
- [11] William Rand. Objective criteria for the evaluation of clustering method [J]. *J.Am. Assoc.*, 1971, 66(336): 846-850.
- [12] Wayne Zachary. An information flow model for conflict and fission in small group [J]. *Journal of Anthropological Research*, 1977, 33(4): 452-473.
- [13] Andrea Lancichinetti, Santo Fortunato: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities [J]. *Physical Review E*, 80, 016118, Jul, 2009.
- [14] Huawei Shen, Xueqi Cheng, Kai Cai, and Maobin Hu. Detect overlapping and hierarchical community structure [J]. *Physica A*, 2008, 388: 1706-1712.
- [15] Roger Guimera, Marta Sales-Pardo, Luís A. Nunes Amaral. Modularity from fluctuations in random

- graphs and complex networks [J]. *Physical Review E*, 70, 025101, Aug. 2004.
- [16] Jordi Duch and Alex Arenas. Community detection in complex networks using extremal optimization [J]. *Physical Review E*, 72, 027104, Jun. 2005.
 - [17] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. Fast unfolding of communities in largenetworks. [J]. *J. Stat. Mech.* Oct. 2008.
 - [18] Yuxin Zhao, Wen Jiang, Shenghong Li, Yinghua Ma, Guiyang Su, Xiang Lin. *Nerocomputing 2013*(Accepted).
 - [19] Jingfei Du, Jianyang Lai, Chuan shi. Multi-objective Optimization for Overlapping Community Detection [J]. *Advanced Data Mining and Applications*, 2013.
 - [20] Chuan Shi, Philio Yu, Zhengyu Yan, Yue Huang, Bai Wang. Comparison and selection of objective functions in multi-objective community detection. [J]. *Computational Intelligence*, 2013.
 - [21] Maoguo Gong, Lijia Ma, Qingfu Zhang, Licheng Jiao. Community detection in networks by using multiobjective evolutionary algorithm with decomposition. [J]. *Physica A*, 2012.
 - [22] Mark Newman and Elizabeth Leicht Mixture models and exploratory analysis in networks. [J]. *Proc. Natl. Acad. Sci.*, 2007, 104(23): 9564-9569.
 - [23] Wei Ren, Guiying Yan, Xiaoping Liao, and Lan Xiao Simple probabilistic algorithm for detecting community structure [J]. *Physical Review E*, 79, 036111, Mar. 2009.
 - [24] David Blei, Andrew Ng and Michael Jordan. Latent Dirichlet Allocation. [J]. *Journal of Machine Learning Research*, Mar. 2003.
 - [25] Le Yu, Bin Wu, and Bai Wang. LBLP: Link-Clustering-Based Approach for Overlapping CommunityDetection. [J]. *TSINGHUA SCIENCE AND TECHNOLOGY*, 2013, 18(4).
 - [26] Youngdo Kim and Hawoong Jeong. Map equation for link communities [J]. *Review E*, 84, 026110, Aug. 2011.
 - [27] Lada Adamic and Natalie Glance. The Political Blogosphere and the 2004 U.S. Election:Divided They Blog. [J]. in *Proceedings of the WWW-2005Workshop on the Weblogging Ecosystem*, 2005.
 - [28] Martin Rosvall, D. Axelsson, and Carl Bergstrom. The map equation. [J]. *Eur. Phys. J.*, 178, 2009.
 - [29] Songchang Jin, Aiping Li, Shuqiang Yang, Wangqun Lin, Bo Deng, Shudong Li. A MapReduce and Information Compression based Social Community StructureMining Method. [J]. *16th International Conference on Computational Science and Engineering*, 2013.
 - [30] Illés Farkas, Dániel Ábel, Gergely Palla. Tamás Vicsek, “Weighted network modules. [J]. *New Journal of Physics*, Jun. 2007.
 - [31] Jukka-Pekka Onnela, Jari Saramäki, János Kertész, Kimmo Kaski. Intensity and coherence of motifs in weighted complex networks [J]. *Physical Review E*, 71, 065103, Jun. 2005.
 - [32] Jussi Kumpula, Mikko Kivelä, Kimmo Kaski, Jari Saramäki. Sequential algorithm for fast clique percolation [J]. *Physical Review E*, 78, 026109, Aug. 2008.
 - [33] Xiaojin Zhu, and Zoubin Ghahramani. Learning from Labeled and Unlabeled Data withLabel Propagation [J]. *Technicalreport, CMU CALD tech report CMU-CALD-02*, 2002.

- [34] Steve Gregory. Finding overlapping communities in networks by label propagation. [J]. New J. Phys. 12, 103018, Oct. 2010.
- [35] Ian Leung, Pan Hui, Pietro Liò, and Jon Crowcroft. Towards real-time community detection in large networks. [J]. Physical Review E, 79, 066107, Jun. 2009.
- [36] Yuxin Zhao, Shenghong Li, Xiuzhen Chen. Community Detection Using LabelPropagation in Entropic Order. [J]. 12th International Conference on Computer and Information Technology, 2012.
- [37] Hao Lou, Shenghong Li, Yuxin Zhao. Detecting community structure using label propagation with weighted coherent neighborhood propinquity. [J]. Physica A, 2013, 392: 3095-3105.
- [38] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. “Defining and identifying communities in networks[J]. Proc. Natl. Acad. Sci, 2004, 101(9): 2658-2663.
- [39] Conrad Lee, Fergal Reid, Aaron McDaid, and Neil Hurley Detecting highly overlapping community structure by greedy clique expansion [J]. In Proc. SNAKDD Workshop, 2010: 33-42.
- [40] Jianbin Huang, Heli Sun, Jiawei Han, Boqin Feng. Density-based shrinkage for revealing hierarchical and overlapping community structure in networks [J]. Physica A 2011: 2167-2171.
- [41] Amanda Traud, Eric Kelsic, Peter Mucha, and Mason Porter. Comparing community structure to characteristics in online collegiate social networks [J]. SIAM Rev, 2011, 53(3): 526-543.

Evolution Analysis of Virtual Communities

4.1 Introduction

In researches on social networks, the research on the structure of virtual communities has attracted wide attention from scientists. Virtual communities which constructs a social network are groups or clusters, and they can reflect the characteristic of gathering locally form the individual behaviors in the network . There are a lot of explicit or implicit virtual community structures of various types in online social networks , such as circles in renren.com, groups in douban.com and so on. These virtual communities' structures are not immutable, but evolve with the evolution of online social network structure over time. Traditional researches on the structure of the virtual communities focus on static network. With the arrival of the era of big data, access and analysis of large-scale evolution data in dynamic network become possible, and it has become a new trend to turn from analyzing static networks to researching, the evolution of dynamic networks in recent social network researches. The evolution of virtual communities is closely related with functions of social networks such as diffusion, invulnerability, cooperation and synchronization, and it also plays a fundamental role in the evolution of the social network. Therefore, the evolution issues of virtual communities in the social network have important research value. This chapter mainly introduces contents related to the evolution of virtual communities. Those contents are organized as follows: Section 4.2 introduces three basic mechanisms for merging of virtual communities, i.e. period closure, preference connection and aging factors; Section 4.3 further analyzes the effects of structural diversity of

individuals from the aspect of accumulative effect by individuals joining virtual communities, and then considers the effects of structural balancing factors on the evolutions of virtual communities; Section 4.4 introduces the detection algorithm for evolving virtual community based on the similarity comparison at adjacent moments, the detection algorithm for evolving virtual community based on evolution clustering analysis, the detection algorithm for evolving virtual community based on Laplacian dynamics, the detection algorithm for evolving virtual community based on clique percolation algorithm, the detection algorithm for evolving virtual community based on trend analysis on node behaviors and other typical dynamic virtual community detection algorithms.

4.2 Merging of Virtual Communities

An important structural characteristic during the merging of virtual communities is clustering phenomenon in network, which exists in many real networks such as social network, World Wide Web, reference network and scientist cooperation network. This section mainly introduces the effects of the period closure, preference connection, aging factors and other mechanisms in the merging of virtual communities on the clustering phenomenon, and introduces related models.

4.2.1 Period Closure in Merging of Virtual Communities

An important topological structural characteristic of social network is that every edge has different weights with different physical meanings, such as closeness of relationship and frequency of interaction. Besides social network, many networks can be depicted by weighted network, such as transportation network, metabolism network and other connectivity-based networks. In those networks, the weight of edge has an important effect on the nature and function of network, such as disease transmission^[1], synchronization dynamics of vibrator^[2] and statistics of die body^[3], while the weight of edge has an important effect on the merging of virtual community in social network.

In social network, we call connections between close friends as strong connections and connections between friends that have distant relationship or meets once in a while as weak connections. Strong connection and weak connection are two important types of edge^[4]. As shown in researches, the topological structure of large-scale social network satisfies weak connection assumption^[4], i.e. strong connections mostly appears inside virtual community

in network and weak connections mostly appears between virtual communities. As shown in researches, topological structure like online social network is formed through the micro evolution of two kinds of social networks, i.e. period closure and focus closure^[5].

Period closure refers to the structure formed as nodes in network tend to establish connection with neighbor of its neighbor in the network, which is the main factor for the formation of virtual community. As shown in experiment, the probability of ternary closure decreases exponentially with the increase of geodesic distance between two nodes. On the contrary, focus closure is independent of geodesic distance, but generated by the common interests or activities of two nodes.

By combining the above two micro evolution mechanisms of social network, Jussi Kumpula et al. proposed the merging model of weighted virtual community^[6] with algorithm mainly comprising the three steps (Algorithm 4-1) as shown in Figure 4-1.

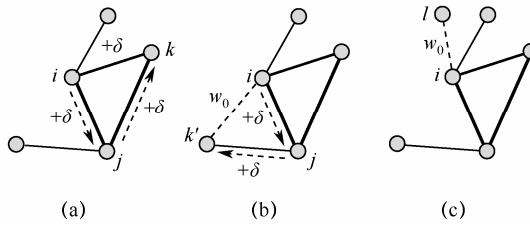


Figure 4-1 Merging algorithm of weighted virtual community [(a) and (b) show local connection mechanism and (c) shows overall connection mechanism (see Reference [6])]

(1) Local connection mechanism: During time interval Δt , each node i connects one of its neighbor node j at the probability of $\omega_{i,j}/s_i$, wherein $\omega_{i,j}$ denotes the weight of i, j connection and $s_i = \sum_j \omega_{i,j}$ denotes the weight of node i . If the selected node j has other neighbor nodes besides i , randomly select a node k from them at probability $\omega_{j,k}/(s_j - \omega_{i,j})$. If there is no other edge between nodes i and k , generate the edge between i and k at the probability of $p_\Delta \Delta t$ with the weight of $\omega_{i,k} = \omega_0$; otherwise, the weight of this edge increase by δ . $\omega_{i,j}$ and $\omega_{j,k}$ increase by δ regardless of whether edge exists between i and k . The above process reflects the period closure mechanism.

(2) Overall connection mechanism: If a node has neighbor nodes, such node connects to a randomly-selected node at the probability of $p_r \Delta t$ to form an edge with weight of ω_0 ; otherwise, such node connects to a randomly-chose node to form an edge with weight of ω_0 . This process is similar to focus closure as a node other than neighbor node of the selected node is selected.

(3) Removal mechanism: Remove all nodes and its connected edges at the probability of $p_d \Delta t$ and replace the removed node with new node to keep the total number of nodes unchanged.

wherein the proportion of p_d and p_r reflect the internal density of the network community generated from this model. The weight of network edge can be adjusted by parameter δ . When $\delta=0$, an unweighted network is generated from this model. When $\delta>0$ increases monotonically, the internal density of the network community generated from this model increases and the internal edge weight of community increases due to local edge mechanism. With the increase of δ , after some edges are selected, the edge weight of its triangles increases rapidly, making edges of these triangles easily to be selected repeatedly, thereby forming community structure with dense connections around those triangles. Figure 4-2 shows the structural graph of network generated from this model through different parameters.

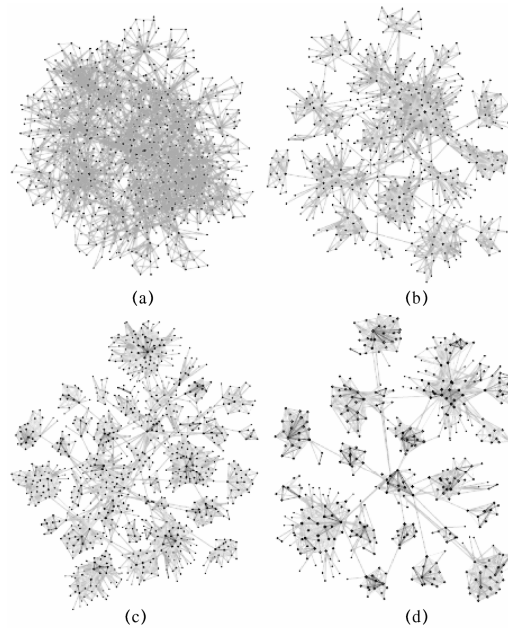


Figure 4-2 (a) $\delta=0$, (b) $\delta=0.1$, (c) $\delta=0.5$, (d) $\delta=1$ network structural graph [edge color from light color (weak connection) to dark color (strong connection) (see Reference [6])]

Network generated from this model has not only community structure with adjustable density, but also some typical characteristics of real social network:

(1) Degree distribution follows exponential distribution with asymmetry.

- (2) Nodes with big degree in network tend to connect other nodes with big degree and represent positive correlation.
- (3) The network has high clustering coefficient of $c(k) \sim 1/k$.
- (4) The network has average diameter of $\log N$ and represent small-world phenomenon.

4.2.2 Preference Connection in Merging of Virtual Communities

Social network, like many other complex network (e.g. metabolism network, computer network), represents small-world phenomenon and follows power-law distribution. Duncan Watts and Steven Strogatz firstly proposed the WS small-world model for modeling network with high aggregation; Albert-László Barabási and Réka Albert firstly proposed the BA model for modeling network with power-law distribution. However, neither WS nor BA model is capable of modeling social network with both high aggregation and power-law distribution. This section mainly introduces two types of model for modeling network with both high aggregation and power-law distribution.

Type I is clustering scale-free network model^[7] proposed by Peter Holme et al. First, we define network $G=(V,E)$, wherein V denotes the set of all nodes in network and E denotes the set of all edges in network. We assume there is no multi-edge between any two nodes and clustering coefficient of the network is $\gamma = \langle \gamma_v \rangle$, wherein $\langle \cdot \rangle$ denotes the average of all node clustering coefficients. The clustering coefficient γ_v of node v is defined as follows: assume the degree of node v is k_v , then the possible number of edges among k_v nodes is $k_v(k_v-1)/2$; define $|\xi(\Gamma_v)|$ as the number of actual edges among those nodes, then $\gamma_v = 2|\xi(\Gamma_v)|/k_v(k_v-1)$. If $\gamma=1$, the network is fully connected and the clustering coefficient of network generated from BA model is $\gamma \approx 0$, thus it has no small-world phenomenon.

The algorithm of clustering scale-free model is as follows (Algorithm 4-2): To generate power-law distribution, first introduce the following BA model mechanism into clustering scale-free model.

- (a) The initial network has m_0 nodes but no edge.
- (b) Add 1 node v in network in each step which connects m edges.
- (c) The probability for connecting each new node v to an existing node w is $P_w = k_w / \sum_v \in V k_v$.
- (d) To generate high aggregation characteristic, the following preference connection mechanism is added to this model.

If node v connects to w in the above step (c), randomly selects a neighbor node of w for connecting to v . If all neighbor nodes of w are connected to v already, return to step (c).

In each loop, the algorithm first carry out step (b) and (c), then carry out step (d) at the probability of P_t and step (c) at the probability of $1-P_t$. The average number of triangles connected to each node is $m_t = (m-1)P_t$, wherein m_t denotes control parameter of this model. When $m_t=0$, the clustering scale-free network model degenerates into BA model as shown in Figure 4-3.

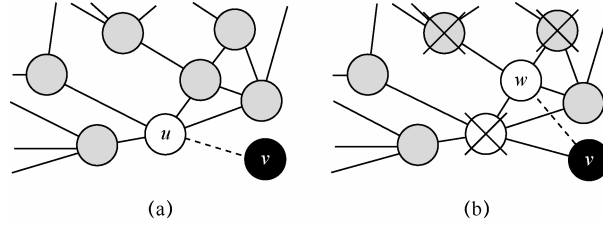


Figure 4-3 (a) In step (c) of clustering scale-free network model, new node v randomly connects to existing node u in network. (b) In step (d), new node v randomly connects to neighbor node w of node u .

× means nodes that cannot connect to v (see Reference [7])

Let's first calculate the degree distribution of network generated from clustering scale-free model. The degree increment of any node v by carrying out step (c) can be represented as

$$\Delta k_v / \Delta t = A k_v / \sum_{w \in I'} k_w$$

wherein A denotes normalization factor and k_v denotes the degree of node v . The degree increment of any node v by carrying out step (d) can be represented as

$$\Delta k_v / \Delta t = \sum_{w \in \Gamma_v} k_w (1/k_w) / \sum_{w \in I'} k_w = k_v / \sum_{w \in I'} k_w$$

wherein Γ_v denotes the set of neighbor nodes of node v . As m_t step (d) and $m-m_t$ step (c) were carried out in each loop of the algorithm of this model,

$$\Delta k_v / \Delta t = m_t (k_v / \sum_{w \in I'} k_w) + (m - m_t) (k_v / \sum_{w \in I'} k_w) = k_v / 2t$$

and further

$$k_v \propto t^{\frac{1}{2}}$$

This result is same to that of BA model. It is easy to obtain degree distribution of the network $P(k) \sim k^{-3}$ as shown in Figure 4-4, indicating that clustering scale-free network model can generate network with power-law distribution, but not scale-free network with

given exponent.

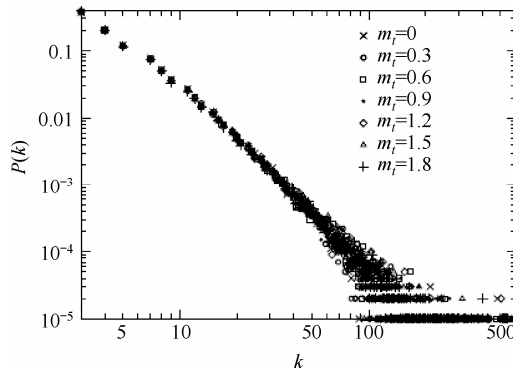


Figure 4-4 Node degree distribution of network generated from clustering scale-free network model.

When number of nodes $N=10^5$, $m=m_0=3$. When $m_t=0$, the degree distribution is similar to that of BA model (see Reference [7])

In addition, the value of parameter m_t in this model affects the proportion of triangles in network and further the clustering coefficient in network. Figure 4-5 (a) shows the function graph of clustering coefficient γ and number of nodes N according to different parameter m_t . We found that, for non-zero m_t , the increase of γ along with N tends to non-zero finite constant. When $m_t=0$, the increase of γ along with N tends to zero. Figure 4-5 (b) shows that clustering coefficient γ increase monotonically along with m_t in approximate linear relation, indicating that this model can generate network with high aggregation with clustering coefficient controlled by parameters. In addition, $1 \sim \log N$ for any $m_t > 0$ indicates that this model can also generate small-world characteristic.

The next is the second type of model for modeling high aggregation and power-law distribution, i.e. acquaintance network^[8] proposed by Joern Davidsen et al. The algorithm of this model is defined as follows (Algorithm 4-3).

(a) Randomly select two neighbor nodes of a reference node and connect the two nodes if they are not connected. If the number of neighbor nodes of the selected node is less than two, randomly select another node for connecting the reference node.

(b) Randomly delete a node and all edges connected to it at the probability of p , then randomly connect it to a node in network.

Repeat the above two steps circularly. Different from clustering scale-free network model, the number of nodes in acquaintance network model remains unchanged. The

existing time of nodes in network is limited due to deletion mechanism (b), thus the whole network can finally reach an equilibrium state. The probability p determines the proportion of step (a) and step (b). In general, one person associates with other persons for several minutes or hours each time, but their relationship exists in social network for many years. Therefore, we only consider situations when $p \ll 1$ in the following discussions.

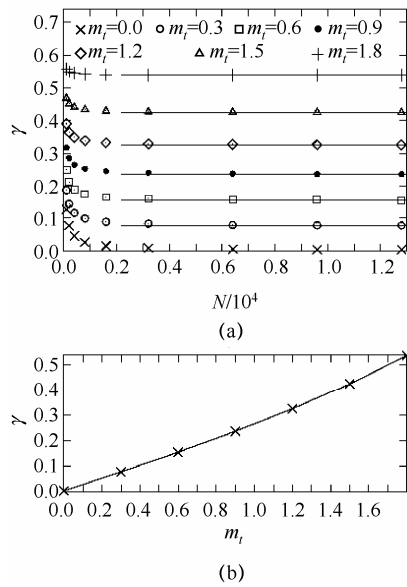


Figure 4-5 Clustering scale-free network model [(a) is the relation graph of clustering coefficient and number of nodes, (b) is the relation graph of clustering coefficient and the average number of triangles connected to each node (see Reference [7])]

Figure 4-6 shows the degree distribution of network at equilibrium state according to different p . As the existing time of each node is finite, node degree has upper limit, thus $p(k)$ only has non-zero value in finite range. As step (a) generates triangle connection and step (b) is Poisson process, when $p \ll 1$, $p(k)$ is determined mainly by step (a) and satisfies power-law distribution; in addition, the range of length for $p(k)$ has non-zero value increases with the decrease of p . For relatively big p , step (b) has bigger effect on degree distribution $p(k)$ and thereby generates similar exponent distribution. When $p \approx 1$, step (b) plays a main role and turns the connecting process into Poisson process. Therefore, acquaintance network model can model social network with power-law distribution and exponential distribution.

For clustering coefficient of model generation network, Table 4-1 gives average degree

$\langle k \rangle$ for different p , secondary moment of degree $\langle k^2 \rangle$, average clustering coefficient C , upper limit C' of coefficient of network with the same degree distribution and without triangle connection, clustering coefficient C_{rand} of random graph with the same number of nodes. For any random network with the same number of nodes and average degree $\langle k \rangle$, if the connecting probability of any two node is $p_{\text{link}} = \langle k \rangle / (N-1)$, clustering coefficient is $C_{\text{rand}} = p_{\text{link}}$. Therefore, for networks with unchanged number of nodes, C_{rand} is directly proportional to average degree of network $\langle k \rangle$ (see Table 4-1). For networks with the same degree distribution and random edges, M. E. J. Newman et al. used the function generation method of random graph and obtained the upper limit^[9] of average degree coefficient of such random network

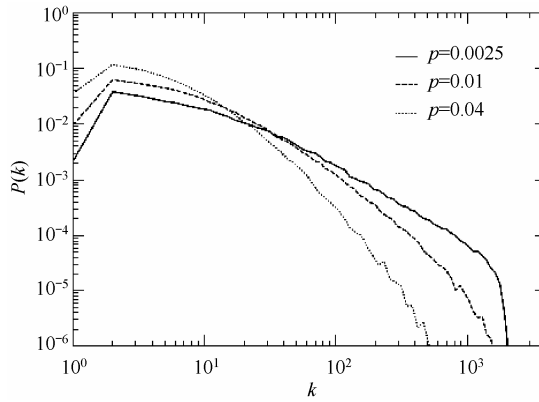


Figure 4-6 Degree distribution of network generated from acquaintance network (see Reference [8])

$$C' = ((\langle k^2 \rangle / \langle k \rangle) - 1)^2 / \langle k \rangle N$$

For networks following Poisson degree distribution, $C' = C_{\text{rand}}$. As shown in Table 4-1, the clustering coefficient of network generated from acquaintance network model is much bigger than that of random graph with the same degree distribution.

Besides, the average path length of model generation network is ($p=0.04$) $l(N) \sim \log N$, i.e. small-world characteristic. Mark Newman et al. used the method to generate function and gave the approximate expression^[9] of average path length l of random graph with any degree distribution:

$$l' \approx \log(N / \langle k \rangle) / \log((\langle k^2 \rangle - \langle k \rangle) / \langle k \rangle) + 1$$

For networks following Poisson degree distribution

$$l_{\text{rand}} \approx \log N / \log \langle k \rangle$$

As shown in Table 4-1, for $p=0.0025$, $l' \approx 1.59$, $l_{\text{rand}} \approx 1.77$, $l' < l_{\text{rand}}$ is the hub node with high degree due to the existence of scale-free network. The value results show the average path length of network generated from acquaintance network model is $l=2.38$, which is another strong evidence for the small-world phenomenon of such network.

Table 4-1 Acquaintance network model^[8]

P	$\langle k \rangle$	$\langle k^2 \rangle$	C	C	C_{rand}
0.04	14.9	912	0.45	0.036	0.0021
0.01	49.1	13,744	0.52	0.29	0.0070
0.0025	149.2	99,436	0.63	0.43	0.021

4.2.3 Aging Factors in Merging of Virtual Communities

This section introduces the generation model of document network clustering phenomenon. Document network mainly includes reference network, online social network, etc., with one node therein denoting an article or a page containing multimedia information. The nature of document network is affected not only by its topological structure, but also the semantic relation between node content. Research on document network is of great importance on page search and information retrieval^[10-12]. This section focuses on the evolution process of document network, and research the effects of node similarity, node degree, aging characteristic of nodes and other factors on aggregation to obtain the generation model of clustering phenomenon of document network.

The most important model for document network research is the degree similarity mixture model^[11] proposed by Filippo Menczer. Filippo Menczer first researched the correlation between document network evolution and content similarity between documents. To research on the similarity between two documents, Filippo Menczer proposed the measurement formula for document content similarity:

$$\sigma_c(d_1, d_2) = \|\bar{d}_1 \cdot \bar{d}_2\| / \|\bar{d}_1\| \cdot \|\bar{d}_2\|$$

wherein \bar{d} is the vector representation of content of d document. By researching the similarity distribution of webpage (DMOZ) and scientific article (PNAS), F. Menczer found that the similarity distribution of content of connected documents is significantly different from that of all documents, and the connection probability between documents increases with the similarity of document content. On this basis, Filippo Menczer proposed the

degree similarity mixture model (DSM model) with the algorithm of model (Algorithm 4-4) as follows.

Add a new node in network each time and the new node has $m=L/N$ edges and network node connection. At step t , the probability for connecting the new node to node i is

$$P_r(i) = \alpha k_i / m_t + (1 - \alpha) P'_r(i)$$

wherein $P'_r(i) \propto (1/\sigma_c(i,t) - 1)^{-\gamma}$, $i < t$, k_i denotes the degree of node i , and parameter is obtained by calculating real data. The first nomial on the right side of the above formula indicates that the node tends to connect a node with big degree in network, which is similar to the preference connection mechanism in BA model. The second nomial on the right side of the above formula indicates that the node tends to connect a node with similar content, with $0 \leq \alpha \leq 1$ as the parameter of preference connection mechanism, which controls the probability for connecting node with big degree and similar content. If $P_r(i) = 1/t$, the new node connects to node in network in a completely random manner, which is called as degree uniformity mixture model (DUM model). Compared with degree uniformity mixture model, degree similarity mixture model can better fit the similarity of content of nodes in network.

Xueqi Cheng et al. discovered the triangle clustering characteristic of document network, and thereby proposed the concept of triangle similarity and degree similarity preference model (DSP model)^[13]. They first defined the connection probability $P(\sigma_c) = M^*(\sigma_c) / M(\sigma_c)$ between two documents with content similarity of σ_c , wherein $M(\sigma_c)$ denotes the number of node pairs with content similarity of σ_c and $M^*(\sigma_c)$ denotes the number of actually connected node pairs in network. As shown by the value results, the connection probability between two documents increases with the similarity of their content. Take PNAS reference network for example, if the similarity between two articles is $\sigma_c = 0.5$, the probability for reference relation between them is $P(\sigma_c) = 50\%$; when $\sigma_c < 0.2$, $P(\sigma_c)$ is very small. They further discovered that two documents similar to the same document are also similar, and proposed triangle similarity for describing such triangle relation of similarity, i.e. $R_{ijk}^\Delta = \min\{\sigma_c(i,j), \sigma_c(i,k), \sigma_c(j,k)\}$ and triangle connection probability $P(R^\Delta)$, wherein $P(R^\Delta)$ denotes the probability that three nodes with triangle similarity of R^Δ form a triangle. The triangle given here is weak triangle, in which any two nodes among the three nodes have at least one directed edge. As shown in value results,

triangle connection probability is sensitive to triangle similarity. For WT10g data, when the triangle similarity increases from 0.1 to 0.5, the triangle connecting probability increases by two magnitudes; for PNAS reference network data, the triangle connecting probability increases by 4 magnitudes. Xueqi Cheng et al. further proposed the DSP model for document network with the algorithm (Algorithm 4-5) as follows.

(a) Expansion process of network: Add a new node in network each time at probability p which is determined by the number of nodes and edges in network, i.e. $p=N/(N+L)$. The average degree of nodes in network is $\langle k \rangle = 2L/N = 2(1-p)/p$.

(b) DSP preference connection process: Two disconnected nodes in network connect at the probability of $1-p$. If this directed edge starts from node i and ends at j , the connection probability between i and j is

$$\Pi(i) = (k_i^{\text{out}} + \beta_1) / \sum m(k_m^{\text{out}} + \beta_1)$$

$$\Pi(j) = (k_j^{\text{in}} + \beta_2) (\sigma_c(i,j) + \alpha) / \sum l[(k_l^{\text{in}} + \beta_2) (\sigma_c(i,l) + \alpha)]$$

wherein k_i^{out} denotes the out-degree of node i and k_j^{in} denotes the in-degree of node j with $i \neq j$, parameters β_1, β_2, α having positive value. Parameters β_1, β_2 can ensure that, at the initial stage of the algorithm, nodes with $k^{\text{in}} = k^{\text{out}} = 0$ can connect to other nodes, while parameter α can ensure the connection between documents with significantly different content. All these conditions can ensure that this model conforms to actual situations.

Compared to DSM model, DSP model can better fit the structural characteristic of document network, such as degree distribution and average clustering coefficient. In addition, DSP model shows great advantages in respect of fitting the function relation of triangle connection probability and triangle similarity.

Fuxin Ren et al. thought that the merging of clustering phenomenon in document network is affected by not only document similarity, document popularity (or the number of document edges), but also aging factors of nodes. On this basis, they proposed degree and age preference connection – clique neighbor preference connection model (DAC model)^[14]. The algorithm of this model (Algorithm 4-6) is mainly composed of two parts.

(a) Degree and age preference connection: The connection probability of a new node i and an existing node j is $\Pi_{ij} \propto k_j^{\text{in}} \times t_j^{-\alpha}$, wherein k_j^{in} denotes the in-degree of node j , $t_j = i - j$ denotes the age of node j , $\alpha > 0$ denotes attenuation parameter. This

probability in the form of power-law is adopted by many models, such as Dorogovtsev–Mendes (DM) model^[26].

(b) Clique neighbor preference connection: All nodes of clique that nodes i and j belong to are connected at the probability of $\beta(0 \leq \beta \leq 1)$. If node j belongs to many cliques at the same time, randomly selects a clique s at the probability of P_s for connecting node i to all nodes in clique s , wherein P_s is directly proportional to the number of node in clique s . If all neighbor nodes of i can connect to i without those nodes in clique, node i can select node connection by degree and age preference connection (a) at the probability of $1-\beta$. Obviously, β controls the increase speed of clustering coefficient in network in a directly proportional manner.

As shown in experiment results, DAC model can well simulate in-degree distribution of reference network, scale distribution of connected components, increasing law of the number of triangles with the number of nodes in network, function relation between average clustering coefficient and node degree. In addition, DAC model can well simulate the relation between edge density and node out-degree in network. Main advantage of this model lies in revealing the relation between time sequence characteristic (aging factors) of node and clustering phenomenon of document network.

4.3 Evolution of Virtual Communities

The evolution process of virtual communities in online social network is very complicated and faces many influencing factors. An important and challenging topic in research on social network is to mine key factor in the evolution of virtual communities. This section mainly uses empirical analysis and mathematical model to introduce the effects of three basic factors of individual user on the evolution of virtual communities, i.e. accumulative effect, structural diversity and structural balance.

4.3.1 Accumulative Effect in Evolution of Virtual Communities

A core problem in social science research is how virtual communities evolve over time and why the scale of virtual communities increases. In the field of digitalized information, MySpace, LiveJournal and other online social networks gradually become the mainstream platform for information network exchange along with the scale's increment of virtual communities, while acquisition and analysis of

large-scale data generated from the evolution process of virtual communities become a difficult problem. To understand the increase process of virtual communities, we will mainly carry out research on the following key problems: Which factor determines a user to join a virtual community? This section introduces research results on those problems by Lars Backstrom et al.^[15]

Two datasets are adopted by Lars Backstrom et al.: friend relationship and virtual community data on LiveJournal as well as cooperators and conference article data on DBLP. To discuss the influencing factor for a user to join a virtual community, they define a user that has friends but has not joined the virtual community as fringe. For a fringe, Lars Backstrom et al. first considered the relation between its probability of joining the virtual community and its number of friends in the virtual community, i.e. whether the possibility for a fringe to join the virtual community has accumulative effect. In LiveJournal, the function relations for the probability for a fringe to join the virtual community $P(k)$ and its number of friends in virtual community k are obtained in the following steps:

- (1) Acquire two snapshots of users included in different virtual communities for a month.
- (2) Construct all ternary groups: (u, C, k) , wherein C denotes virtual community and u doesn't belong to C in the first snapshot. At the time of the first snapshot, u has k friends in virtual community C .
- (3) Given a k , $P(k)$ denotes the proportion that u belongs to the ternary group (u, C, k) of C in the second snapshot.

The function relation of $P(k)$ and k in DBLP network are obtained in a similar manner, except it acquires snapshots in a year and considers the proportion for a user to “join” a meeting in a year. Figure 4-7 and Figure 4-8 respectively give the function relation for the probability for a user to join the virtual community $P(k)$ and its number of friends in virtual community k in LiveJournal and DBLP network. The two functions have similar curves and both have the “diminishing returns”, i.e. monotonically increase curve with slower and slower increment speed, which is totally different from “S Type” logistic function curve obtained in many propagation models. In logistic curve, $P(k)$ increases slowly with relatively small k ; when k is around certain median, $P(k)$ increases faster; when k is relatively large, $P(k)$ increases slowly again. The best fitting function of this curve is $P(k)=a\log k+b$ with parameters a and b .

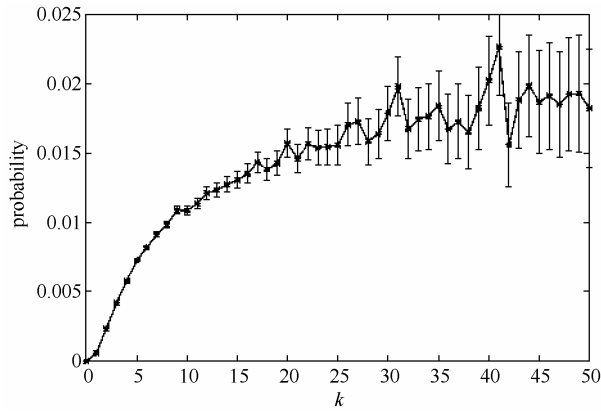


Figure 4-7 The relation graph of the probability for an user to join the LiveJournal virtual community and its number of friends in such virtual community (see Reference [15])

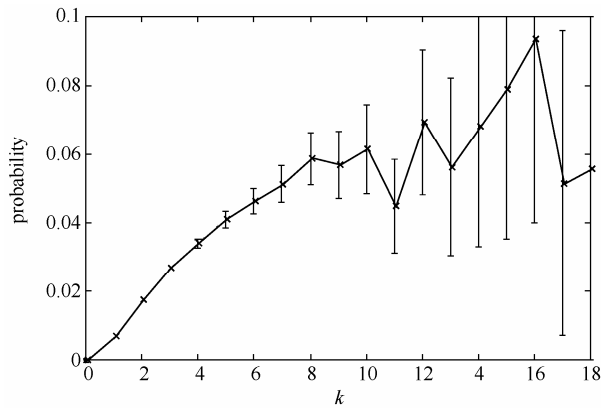


Figure 4-8 The relation graph of the probability for an user to join the DBLP virtual community and its number of friends in such virtual community (see Reference [15])

As shown in experiment, the top layer of decision-making tree of LiveJournal and DBLP are very stable in many experiments and the connection situations of an user with internal friends in virtual community play an important role in prediction. The next problem to be discussed is what's the relation between the connection density degree with internal friends in virtual community of a fringe and its joining in such community, which can be expressed as follows: For a fringe user u with the number of edges $e(S)$ to friends in virtual community S , define the connection intensity between friends as $\varphi(S) = 2e(S) / (|S|(|S|-1))$, which indicates the proportion that a fringe user connects to all its friends in virtual community S , wherein the possible

number of edges between $|S|$ friend is $|S|(|S|-1)/2$. As shown in the results, the fringe tends to join the virtual community when $e(S)$ and $\phi(S)$ are large. For fixed $k=3, 4, 5$, Figure 4-9 shows the function relation between the probability of user joining the virtual community and connection intensity $\phi(S)$, wherein an user tends to join the virtual community when the connection intensity between friends in the virtual community is large. According to social capital argument^[16], in the same virtual community, friends knowing each other is more reliable than strangers, which indicates that an user will be supported by richer local social structure when it joins the virtual community. Therefore, the social capital argument may be the important theoretical support for the conclusion in Figure 4-9. In addition, weak connection assumption indicates that users expect to make friends with strangers in the same virtual community^[17] with a possible reason that users in LiveJournal consider reliability more important than information. Lars Backstrom et al. further inspected the function relation between the expansion speed of virtual community and triangle density therein. As shown in Figure 4-10, expansion speed of virtual community is slow when density is high, which is strange. A possible explanation for this phenomenon: High triangle density indicates that new members in virtual community ceased to increase at certain time in the past and only internal members connected to each other thereafter.

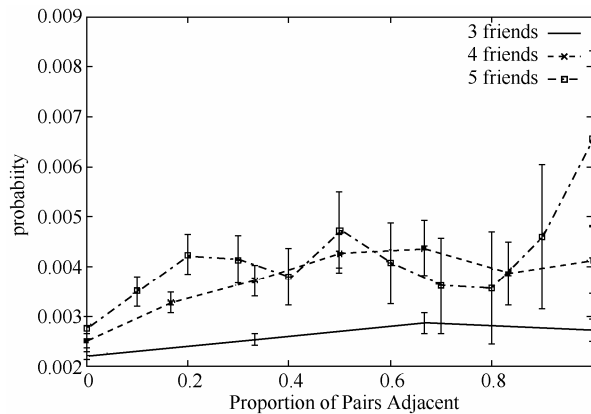


Figure 4-9 Function relation between the probability of an user joining LiveJournal virtual community and its connection degree with friends in such community (see Reference [15])

This section mainly discusses the evolution process of virtual community in social network overtime, including the behavior of individual joining virtual community and the overall increasing behavior of scale of virtual community. As shown in the experiment results, individual joining virtual community shows accumulative effect during the

evolution process of virtual community. The next two sections will further discuss more and more complex influencing factors in the evolution process of virtual community.

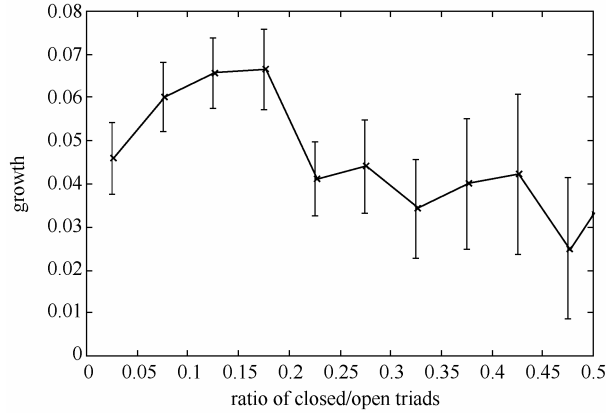


Figure 4-10 Function relation between the scale increasing speed of virtual community and ternary closure density in virtual community (see Reference [15])

4.3.2 Structural Diversity in Evolution of Virtual Communities

Propagation process of information, virus, behavior, etc. in social network relies on the topological structure of network at different level. This section introduces research results^[18] by Johan Ugander et al. on the effect of social network structure on user behavior. Two kinds of user behaviors on Facebook is considered in this section, i.e. recruitment and engagement process; the former refers to the process that an user receives invitation email from a Facebook user and complete recruitment on Facebook; the latter refers to the process that an user engage in specific activities after such recruitment. Though both carried out on Facebook, the two processes have significant difference in specific process, and Johan Ugander et al. mainly considered the effect of structural diversity of user nodes on the two behaviors. Structural diversity of user nodes refers to the number of user nodes of connected components which neighbor nodes in social network belong to.

Let's first analyze the recruitment process on Facebook. Assume an user A is not the user of Facebook, then users input A's email address into Facebook, so all Facebook users that A may know. Define Facebook users that has A's email address as user A's contacted neighbor users on Facebook who are subset of A's potential friends in the future (see Figure 4-11). In fact, A may know more people, but we cannot use all its friends as the sample to

predict if an user will be recruited to Facebook (recruitment process) as some of them are not recruited. The recruitment process of an user on Facebook is as follows. As Facebook allows its users to send email to their friends for recruitment, which contains not only the inviter's name but also the list of all contacted users. Johan Ugander et al. researched a basic problem by analyzing data including 54 million invitation emails: What's the relation between the probability of user recruitment on Facebook and the structure of its contacted neighbors? A traditional assumption is that the probability monotonically increase with the number of contacted neighbors, while the results given by Johan Ugander et al. are that the probability of user recruitment on Facebook is only related to the connected components comprised of contacted neighbors. Figure 4-12 (a) indicates the relation between Facebook user conversion ratio and connected edge density of contacted neighbors when there is only one contacted neighbor, and the results are there is no significant relation between the two. Figure 4-12 (b) shows that bigger number of contacted neighbors means lower Facebook user conversion ratio under fixed number of connected components comprised of contacted neighbors. In fact, the influencing factor for Facebook user conversion ratio is neither the number of inviters nor edges between them, but the number of connected components comprised of inviters, i.e. structural diversity. As connected component where each contacted neighbor locates can be deemed as different social environments, the number of different social environments in Facebook determine the probability for user recruitment on Facebook. There is an implicit social relation exists on Facebook, i.e. as figures posting photos on Facebook will be labeled, two Facebook users labeled in the same photo have social relation though they have no connection. Figure 4-12 (c) shows that user conversion ratio is lower if two contacted neighbor nodes are labeled more times (more intense social relation) on the same photo regardless of the connection between them. Therefore, social relation intensity is the extension of social relation to a certain degree, and two users with more intense social relation have more similar social environment. More different social relation brings higher user conversion ratio, which explains the reason for lower user conversion ratio brought by higher social relation intensity between contacted neighbors. Finally, consider the effect of inviters' location in the topological structure of all contacted neighbors on user conversion ratio. Figure 4-13 shows the function relation between user conversion ratio and the topological structure of composition graph of contacted neighbors as well as inviters' location. As shown in results, the inviters' location has no significant effect of user conversion ratio, while acceptance probability of recruitment sent by inviters with higher node degree is slightly larger than that with lower node degree.

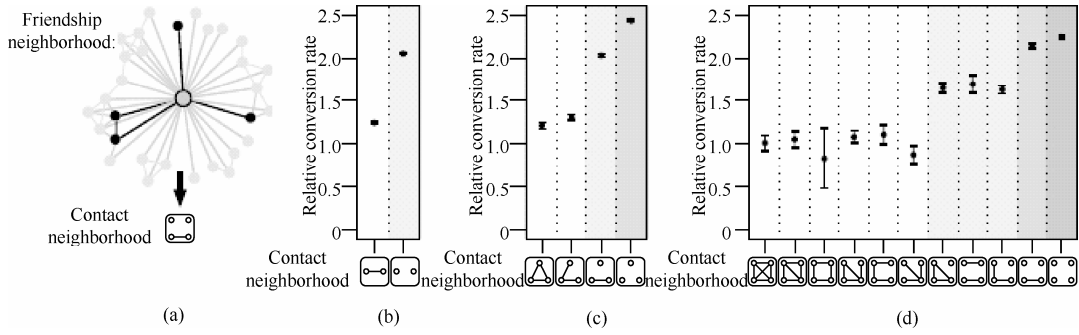


Figure 4-11 (a) Structural graph of contacted neighbors during recruitment process with nodes in light color denoting user's friends and nodes in dark color denoting user's contacted neighbors. Contacted neighbors comprise three connected components. (b~d) relative conversion ratio corresponds to two, three and four contacted neighbor graph (see Reference [18])

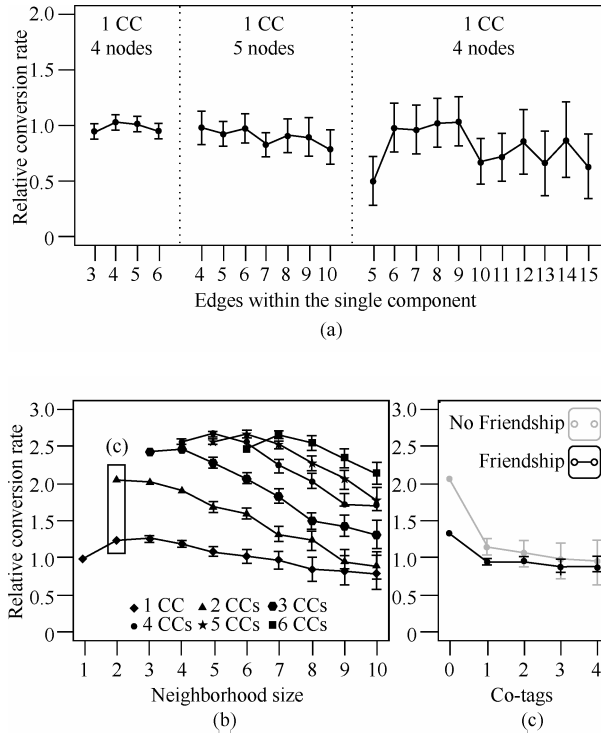


Figure 4-12 During recruitment process, (a) function relation graph between Facebook user conversion ratio and the number of edges in connected components when only one connected component comprised of contacted neighbors, (b) function relation graph between user conversion ratio and the number contacted neighbors when the number of connected components comprised of contacted neighbors is fixed, (c) function relation graph between user conversion ratio and the labeling times of two contacted neighbors on the same photo (see Reference [18])

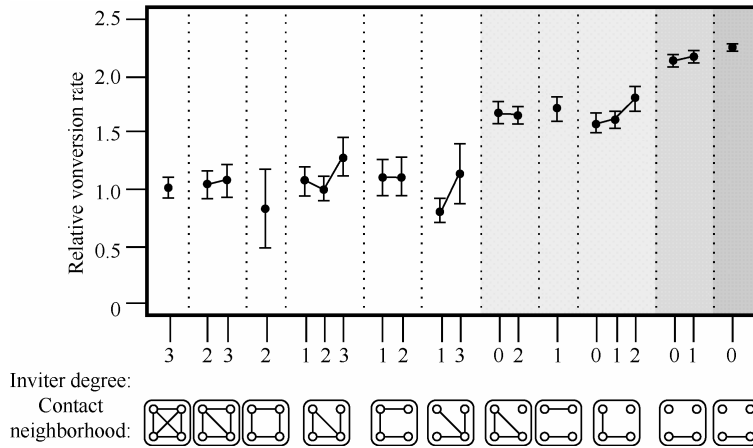


Figure 4-13 During recruitment process, function relation graph between user relative conversion ratio and topological structure of graph comprised of contacted neighbors as well as location of inviter, wherein 4 contacted neighbors exist and the location of inviter is denoted by the node degree in graph comprised of contacted neighbors. Relative conversion ratio is the ratio of actual conversion ratio or user conversion ratio when user has only one neighbor (see Reference [18])

During the login process, the effects of mainly considered. Specifically, we consider whether the structural diversity in the next week after users' recruitment on Facebook in 2010 can be used to predict these users' frequent login three months after recruitment. The standard for frequent user login is Facebook login at least in 6 days out of each week. Friend scale of Facebook users is much larger than that of email. Ten million users recruited to Facebook in 2010 and the number of their friends are 10~50. In addition, as large proportion of connected components comprised of users' friends are single users (nodes), it is not accurate to reflect social environment diversity by the number of connected components comprised of users' friends. To reflect the diversity of social environment more accurately, we give three types of number of induced connected components. Type I number of induced connected components is the number of connected components with node number of k . Type II number of induced connected components is the number of connected components in k -core structure of neighbor nodes. Type III number of induced connected components is the number of connected components in k -brace structure of neighbor nodes. Wherein, embeddedness of edge is defined as the number of common neighbor nodes between two nodes of such edge, and the k -brace structure of a graph is defined as the sub graph after repeatedly deleting edges with embeddedness smaller

than k and isolated nodes. Figure 4-14 (a) and (b) show the example of three types of parameters for measuring social environment diversity of users. Figure 4-14 (c), (d), (e) and (f) show the function relation between user Facebook login frequency and the number of induced connected components. As shown in the results, the larger number of the above three types of induced connected components one week after user recruitment, the higher user login frequency after 3 months. Therefore, the number of induced connected components can well reflect the diversity of social environment that users belong to and effectively predict user login frequency. Figure 4-15 shows the relation between edge density of neighbor nodes of users and user login frequency. As shown in the results, user login frequency increases then decrease along with the increase of edge density, i.e. a peak value exists in the range of (0,1). A possible explanation for the existence of peak value is that too small edge density means lack of social environments and too large edge density means lack of diversity of social environments.

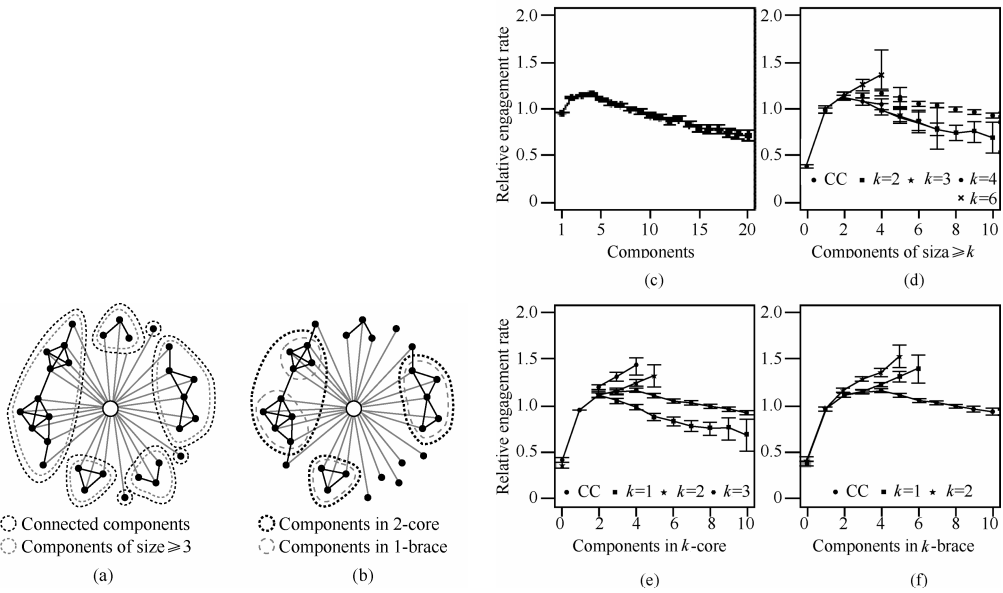


Figure 4-14 (a) Structural graph of all connected components and connected components with scale ≥ 3 of users' neighbors. (b) Structural graph of connected components of users' neighbors 2-core and 1-brace. (c) ~ (e) Relation graph of user relative login frequency of 50 neighbors and induced structure diversity.

Relative login frequency is the ratio of actual login frequency and average user login frequency with 50 neighbors (see Reference [18])

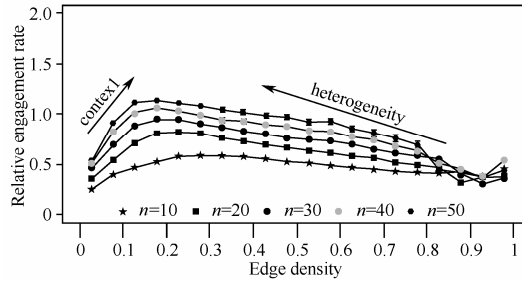


Figure 4-15 When the number of neighbors of users is respectively 10, 20, 30, 40 and 50, the function relation graph of user relative login frequency and internal connection density of neighbors (see Reference [18])

4.3.3 Structural Balance in Evolution of Virtual Communities

Social relation in social network can be classified into two types: positive relationship and negative relationship; the former refers to friendship between an user and another user or support, agreement and other positive emotional factors of an user to another user; the latter refers to the hostile relationship between an user and another user or opposition and distrust of an user to another user. Such positive and negative relationship in social network exists generally and plays an important role in environment, structure, evolution and other aspects of social network. For example, Wikipedia users can give affirmative vote or negative vote on administrator candidates; Epinions users can declare their trust or distrust on certain person; Slashdot users can declare certain person as their friends or enemies. If we denote edge of positive relationship as positive sign and edge of negative relationship as negative sign, a basic problem is that how the symbols of whole network or local network affect the symbol of an edge^[19]. The meaning for researching this problem is to understand the basic principle of generating negative relationship in social network.

Daniel Gruhl et al. first proposed the symbol prediction problem^[20] of single edge, i.e. how to predict the symbol of an edge with known topological structure of whole directed network and known symbols of all edges except for one edge. Formalized definition of such problem is as follows: the symbol of directed edge (x,y) from node x to node y in directed graph $G=(V,E)$, $s(x,y)$, i.e. if (x,y) has positive symbol, $s(x,y)=1$; if (x,y) has negative

symbol, $s(x,y)=-1$; if no edge exists between x and y , $s(x,y)=0$. In some situations, we only care the symbol of directed edge and ignore its direction. If an directed edge (x,y) or (y,x) between node x and y has positive symbol while the directed edge between x and y in another direction doesn't exist or has positive symbol, then $s(x,y)=1$. In a similar way, If an directed edge (x,y) or (y,x) between node x and y has negative symbol while the directed edge between x and y in another direction doesn't exist or has negative symbol, then $\bar{s}(x,y)=-1$. The remaining situations except for the above two is denoted as $\bar{s}(x,y)=0$ (including the situation that (x,y) and (y,x) have opposite symbols). The next task is to predict $s(x,y)$ or $\bar{s}(x,y)$ with all known symbols of edges except for one directed edge (x,y) . Machine learning is used in prediction method and two types of feature vector are used in prediction process. Type I feature vector has 7 features, i.e. in-edge $d_{in}^+(v)$ with all symbols of node v as positive, in-edge $d_{in}^-(v)$ with all symbols of node v as negative, out-edge $d_{out}^+(u)$ with all symbols of node u as positive, out-edge $d_{out}^-(u)$ with all symbols of node u as negative, $d_{out}^+(u)+d_{out}^-(u)$, $d_{in}^+(v)+d_{in}^-(v)$, and the number of common neighbors of u and v (or embeddedness) $C(u,v)$. Type II feature vector considers all triangles containing (u,v) , which construct a 16-dimension vector as there are $2 \times 2 \times 2 \times 2 = 16$ possible different triangles for different symbols and directions. Classify the vector by Logistic regression model with probability expression as follows

$$P(+|x) = \frac{1}{1 + e^{-(b_0 + \sum_i^n b_i x_i)}}$$

wherein (x_1, x_2, \dots, x_n) denotes feature vector and b_0, b_1, \dots, b_n denotes coefficient obtained by learning training data. As shown in experiment results, for two types of datasets, Logistic regression model can achieve good prediction by three types of feature vector (Type I feature vector, Type II feature vector, combination of Type I and Type II feature vector). As the three types of feature vector is only related to local topological structure, local topological structure characteristic is available for effective prediction of edge symbol.

Jure Leskovec et al. further carried out symbol prediction by traditional structural balance theory^[19], wherein some social network relations are consider to be more common and stable than other social network relations^[21,22]. Structural balance theory mainly researches the friendly or opposite relation between three persons and it is more common and stable to consider my enemy's friend as my enemy than my friend's enemy as my enemy, i.e. if triangle relationship forms between w and edge (u,v) , the number of friend

relationship in triangle relationship must be an odd number. In other words, a function can be defined as follows:

$$f_{\text{balance}} : \{\text{types}\tau\} \rightarrow \{+1, -1, 0\}$$

wherein τ denotes triangle relationship (w, u, v) with $f_{\text{balance}}(\tau) = \overline{s}(u, w) \overline{s}(v, w)$. As shown in the results, the Logistic regression model combined with structural balance theory can better carry out symbol prediction on Epinions and Slashdot.

4.4 Detection of Evolving Virtual Communities

We introduced virtual community detection algorithm in static network in Chapter 3. However, online social network changes dynamically over time, and events such as merging, disintegration, merging and splitting cause dynamic evolution of social community structure. Therefore, it is an important problem in virtual community research to identify dynamically-evolving virtual community over time. In general, the task of detection of evolving virtual community is to confirm overall virtual community partition at all times or possible structural form of certain virtual community at the next moment, thereby finally identifying all evolving virtual community sequence in dynamic network by detection method for evolving virtual community. This section mainly introduces several types of basic detection algorithms for evolving virtual community in combination with general reference^[23].

4.4.1 Detection of Evolving Virtual Community Based on Direct Similarity Comparison at Adjacent Moments

Detection method for evolving virtual community based on direct similarity comparison at adjacent moments is the most direct method for confirming evolving virtual community sequence in dynamic network. The thought in this algorithm is as follows: First, confirm the community partition of network at adjacent moments (respectively t moment and $t+1$ moment) by virtual community detection algorithm in static network. Second, compare all detected communities in network at adjacent moments to confirm community C_t in t moment network satisfying certain similar conditions with community C_{t+1} detected at $t+1$ moment and add C_{t+1} to the evolving virtual community sequence which C_t belongs to.

John Hopcroft et al. first researched the detection method for evolving virtual

community^[24] in dynamic reference network by the detection algorithm for evolving virtual community based on direct similarity comparison at adjacent moments. First, carry out community detection on network snapshot in dynamic network sequence by hierarchical clustering algorithm with main thought as follows: Carry out merging from communities with the highest similarity until all elements are included in one community. The following including angle cosine is used to define similarity between nodes and distance between communities:

$$\text{similarity}(i, j) = \cos(r_i, r_j) = \frac{r_i \cdot r_j}{\|r_i\| \|r_j\|}$$

$$\text{dis}(C, C') = \sqrt{\frac{n_C n_{C'}}{n_C + n_{C'}}} (1 - \cos(r_C, r_{C'}))$$

wherein n_C denotes node scale of community C , r_i denotes property vector comprised of all references of network node (article) i , and r_C denotes the normalization sum of all node property vectors in the community. To discover a stable community structure not affected by change of disturbance data, we delete a small part of nodes and edges in the tree diagram obtained in hierarchical clustering process, and define communities affected slightly as natural community. For natural communities in each network snapshot confirmed by hierarchical clustering process, we define matching degree between communities to discover the best matching natural community at adjacent moments, thereby obtaining sequence of evolving virtual community. The calculating formula of matching degree between communities is as follows:

$$\text{match}(C, C') = \min\left(\frac{|C \cap C'|}{|C|}, \frac{|C \cap C'|}{|C'|}\right)$$

The defect of above analysis method mainly appears during implementation of hierarchical clustering algorithm. As its results is unstable, few times of clustering experiment may fail to select valuable communities from hierarchical tree, thus multiple times of clustering experiments is needed. Besides, this method is mainly designed for reference network, and may not be available for definitions in application of other types of dynamic networks, such as node similarity and distance between communities. Reference [25] gives a more general description of method for those defects. Based on the evolution analysis method in Reference [25], Reference [26] give a more logic definition of evolution

events in various communities.

In general, this kind of method is intuitive and easy to operate, but has some problems in accuracy, i.e. incorrect results^[27]. For example, for two communities A_t and B_t in given network with no overlapped nodes, when A_t expands to sufficient scale, A_{t+1} may have overlapped nodes with B_{t+1} (while they belong to different communities in detection algorithm of static virtual community). Such overlap between communities may make the similarity between A_{t+1} and B_t higher than that between A_{t+1} and A_t . Therefore, according to optimal similarity principle, A_{t+1} is included in the sequence of evolving virtual community where B_t belongs to, which leads to incorrect results in members of evolving virtual community.

4.4.2 Detection of Evolving Virtual Community Based on Evolution Clustering Analysis

Detection of evolving virtual community based on direct similarity comparison at adjacent moments mainly carry out independent community detection on network snapshot at adjacent moments to obtain evolution sequence of virtual community. However, this method may cause significant changes of community structures at adjacent moments in the evolution sequence of virtual community. It is possible that such significant change is not caused by dynamic evolution of network. For this defect, based on the clustering method of static network community detection, Reference [28] proposed evolution clustering analysis frame which can carry out detection of evolving virtual community on dynamic network sequence with the same scale. The basic principle of evolution clustering is to determine the community partition of current network according to the structure of current network and community partition of previous network. This method ensures both good partition of current network and little difference from the community partition of the previous network. References [29,30] rephrases spectral analysis clustering technology under the evolution clustering analysis frame. Based on evolution clustering, Yu-Ru Lin et al. introduced a analysis frame which allows a node belongs to multiple communities at the same time to partition dynamic community evolving over time^[31]. This model can be used to discover the best community partition results to fit observation data and time sequence evolution.

This model mainly considers two kinds of factors when evaluating community

partition quality, i.e. snapshot cost (SC) and temporal cost (TC). SC is defined as Kullback-Leibler divergence of similarity matrix of nodes of network at certain moment and community partition results at this moment, and bigger SC brings worse community partition. TC is defined as Kullback-Leibler divergence of community partition results of network at adjacent moment, and bigger TC brings bigger fluctuation of social partition results at adjacent moments. We optimize the following target function to discover the optimal community partition:

$$\text{cost} = \alpha \cdot \text{SC} + (1 - \alpha) \cdot \text{TC}$$

wherein α is the parameter for balancing the two factors.

In general, the analysis of evolving community based on evolution clustering technology combines the detection of optimal community partition of network with member selection of sequence of evolving community. The sequence of evolving community results obtained ensures both close connection between communities at adjacent moments in the sequence and accuracy of network partition. However, the results obtained is evolution sequence of overall structure of network community, not specific evolution analysis of certain community.

4.4.3 Detection of Evolving Virtual Community Based on Laplacian Dynamics

Social network usually has multiple types of nodes or edges due to its complex structure, i.e. heterogeneous network. Reference [32] first researches partition method for network communities with different types of edges. Reference [33] mainly focuses on the stability of communities affected by Laplacian dynamics factors. On this basis, Reference [34] proposed a method available for detecting multislice network community with heterogeneous structure. This method considers intracommunity connection (adjacent matrix) and interslice coupling to propose quality evaluation function similar to modularity and is available for multislice network. “Slice” in multislice network may be network structure in the same network at different moments or the slices of multiple networks in a combined network with different types of edges. In multislice network, node sets are fixed and edges include intracommunity connection and interslice coupling. Multislice network is as shown in Figure 4-16.

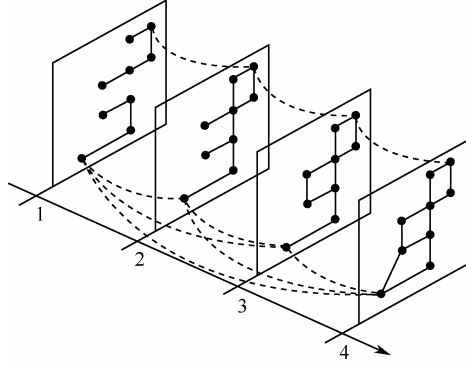


Figure 4-16 Graph of multislice network. There are 4 slices in this network with full line denoting edges between intracommunity nodes and dotted line denoting coupling of interslice nodes (see Reference [34]).

We define community as a node set under certain time scale in which node randomly walk but cannot walk out. For the construction of multislice network Null model, we consider the possibility of reaching node i in s slice from node j in r slice when node j reach equilibrium state, which is discussed in the following two situations:

- (1) i, j at the same slice (adjacent);
- (2) i, j at different slices (coupling).

Finally, we consider the intraslice and interslice movement respectively and obtain corresponding null model:

$$\rho_{is|jr} P_{jr}^* = \left[\delta_{rs} \frac{k_{is}}{2m_s} \frac{k_{jr}}{k_{jr}} + \delta_{ij} \frac{C_{jrs}}{c_{jr}} \frac{c_{jr}}{k_{jr}} \right] P_{jr}^*$$

Apply equilibrium nature of Markov random walk proposed by Renaud Lambiotte et al. in 2008:

$$\begin{aligned} Q_{\text{multislice}}(t) &= \sum_{i,j \in c} \left[\left(e^{t(Q-1)} \right)_{ij} P_j^* - P_i^* P_j^* \right] \\ &= \sum_{i,j \in c} \left[\left(e^{t(Q-1)} \right)_{ij} P_{jr}^* - \rho_{is|jr} P_{jr}^* \right] \\ &= \sum_{i,j \in c} \left[\left(\frac{A_{ijs} \delta_{rs} + C_{jrs} \delta_{ij}}{k_{jr}} \right) \frac{k_{jr}}{2u} - \gamma \left(\delta_{rs} \frac{k_{is}}{2m_s} \frac{k_{jr}}{k_{jr}} + \delta_{ij} \frac{C_{jrs}}{c_{jr}} \frac{c_{jr}}{k_{jr}} \right) \frac{k_{jr}}{2u} \right] \\ &= \frac{1}{2u} \sum_{i,j \in c} \left[\left(A_{ijs} - \gamma \frac{k_{is} k_{jr}}{2m_s} \right) \delta_{rs} - C_{jrs} \delta_{ij} \right] \end{aligned}$$

wherein γ is resolution parameter related to time factor; A_{ijs} is the edge between intraslice node i and j ; C_{jrs} denotes interslice coupling connecting node j in slice r and slice s ; $k_{js} = \sum_i A_{ijs}$ denotes degree of node j in slice s ; $c_{js} = \sum_r C_{jrs}$ denotes coupling intensity between other slices and node j in slice s ; $k_{js} = k_{js} + c_{js}$ denotes connecting intensity between nodes in multislice network. Use the above as target function for optimization to discover the optimal community partition thereunder.

Laplacian dynamics-based research methods on multislice have wider application space than those methods on static network, and allow community partition and quantification of partition quality of network under multiple time dimensions, multiple resolution parameter values and with multiple types of edges.

4.4.4 Detection of Evolving Virtual Community Based on Clique Percolation Algorithm

Another analysis thought for detection of evolving community comes from a new detection method of evolving community proposed in Reference [27] based on clique percolation (CPM) algorithm. A original intention of this method is to solve the member misjudgment problem for sequence of evolving community due to intercommunity overlap in the above direct comparison method. When considering maximum community similarity, this method combines networks at adjacent moments before disintegration instead of directly comparing network community at adjacent moments. The method in Reference [27] uses clique percolation algorithm, which is an important method for static network community detection available for analyzing community structure with overlap. CPM algorithm thinks that community is comprised of a series of mutually-reachable k -cliques (full subgraph with scale of k). It combines adjacent k -cliques to detect communities and nodes within multiple communities are the overlap elements between communities. Dynamic nature of network over time is revealed by discovering the relation between self-correlation function $C_A(t) = \frac{|A(t_0) \cap A(t_0 + t)|}{|A(t_0) \cup A(t_0 + t)|}$ and evolution of virtual community

over time, with relatively obvious network change in relatively large scale. This overlap degree can be calculated through community structures at different moments in composite graph by CPM in Figure 4-17.

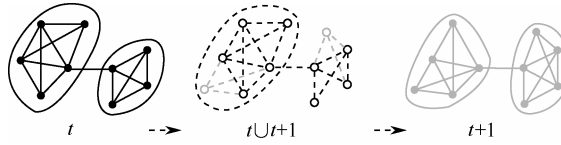


Figure 4-17 Graph of calculating method for overlap degree in community evolution process (see Reference [27])

As CPM algorithm can identify existing overlapping communities, the above evolved community detection method avoids misjudgment in such situations and can identify member of evolved community sequence more accurately. However, this method has a defect, i.e. communities in network can be detected only by CPM algorithm during its implementation. Therefore, cross validation should be carried out on the results of evolved community sequence by other community partition technologies to improve the reliability of results.

4.4.5 Detection of Evolving Virtual Community Based on Trend Analysis on Node Behavior

Detection algorithm for evolving virtual community based on trend analysis on node behavior attribute dynamic evolution of virtual communities to node behavior, and analyze the effects of node behavior on network to determine possible evolution of virtual communities. As shown in Reference [24], enlarged virtual community with dynamic changes of members has longer life period; on the contrary, small-scale virtual community with relatively stable community members has longer life period. Finally, community life period is predicted by analyzing the relation between internal and external members in the virtual community formed in network.

In Reference [24], connected edges between nodes in community can be classified into two types: edges between such node and external nodes (intercommunity edge) and edges between such node and internal nodes (intracommunity edge). To inspect relations between these variables, Reference [24] give a quantified function $W_{out} / (W_{in} + W_{out})$, to denote the proportion of edges between internal node and external node and edges between internal nodes, wherein W_{out} is the weight sum of all intercommunity edges of all member nodes in community and W_{in} is the weight sum of all intracommunity edges of all member nodes in community, When such proportion increases to a certain degree, it is possible for this node to

leave current community and cause dynamic changes. When all nodes tend to leave current community, it may well disintegrate and vanish, i.e. correlation between community and its internal members and external members or communities can indicate community evolution trend.

In general, the above researches doesn't focus on prediction of evolving virtual community but on analysis of node evolution behavior. However, dynamic change of network and virtual communities is essentially caused by node evolution behavior, e.g. network (virtual community) structure change caused by node's exit from network (virtual community). Therefore, possible changes of virtual community structures can be predicted by behavior characteristics related to virtual communities. It is predictable that the thought of confirming possible evolved structure in network by analyzing node behavior trend will facilitate the promotion of new detection method for evolving virtual community.

4.5 Summary

In recent years, online social networking sites, microblogging and other interactive Internet services framed over the Internet have gradually become the mainstream of information network application, and influential sits at home and abroad, such as Facebook, Twitter, Sina Weibo, renren.com and so on, have become important platforms for people's social activities. Virtual community structure is an important structural characteristic of online social networks. Evolution of the virtual community of social network is closely associated with the network's functions, greatly affecting the propagation mode and law of information in social networks, reflecting the characteristics and laws of human activity on social networks. Therefore, the evolution of the virtual communities has important research value and application prospects. This section introduces the research results of the evolution of virtual communities from three aspects: the formation and emerging mechanisms of virtual communities, influencing factors on the evolution of virtual communities , and the detection algorithm of the evolution of virtual communities.

Research on static virtual community has developed for nearly a decade and generated relatively mature research results. However, research on the dynamically-evolved virtual communities in network still wanders at initial stage. With the improvement of data acquisition and analysis technologies, the evolution and analysis problems of virtual

communities will draw increasing attention from researchers. We believe that future research emphasis will be mainly on the following aspects:

(1) In respect of detection algorithm for evolving virtual community, a key problem is how to construct reference graph with general applicability. Currently, there are various types of reference graphs for static network but few for dynamically-evolving virtual community in network, and the existing reference graph has very narrow application scope. Therefore, it is an important problem to perfect existing theoretical model and thereby discover and construct effective dynamic reference graph

(2) Current researches focus on proposing different detection algorithm for evolving virtual community. It is still an problem to select a more effective detection method for evolving virtual community according to different types of networks. For this purpose, researchers are required to carry out comprehensive analysis on complexity, effectivity, comparison and selection, etc., which is an important research direction in the future.

(3) Is it possible to explore the internal mechanism of merging and evolution by tracing and analyzing evolution behavior of important virtual communities in network based on identifying and obtaining evolving virtual community sequence in dynamically-evolved network? This is also an important direction of merging and evolution mechanism research on virtual community in the future.

(4) Currently, there are rich research results for effect of network structure on information propagation but few for effect of information propagation on evolution of virtual community structure. As virtual community constitutes an important part of network structure, it is an important and valuable problem to inspect the effect of information propagation in social networks on the evolution of virtual community structure. With the coming of big data era, these researches will also develop rapidly.

References

- [1] Vittoria Colizza, Alain Barrat, et al. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 2006, 103 (7) : 2015-2020.
- [2] Mario di Bernardo, Franco Garofalo, et al. Synchronizability and synchronization dynamics of weighed and unweighed scale free networks with degree mixing , 2007.
- [3] Jukka-Pekka Onnela, Jari Saramäki, et al . Intensity and coherence of motifs in weighted complex

- networks[J]. Physical Review E, 2005, 71 (6) : 065103.
- [4] Mark Granovetter . The strength of weak ties[J]. American journal of sociology, 1973, 78 (6) : 1360.
- [5] Gueorgi Kossinets, Duncan James Watts . Empirical analysis of an evolving social network[J]. Science, 2006, 311 (5757) : 88-90.
- [6] Jussi Kumpula, Jukka-Pekka Onnela, et al. Emergence of communities in weighted networks[J]. Physical review letters, 2007, 99 (22) : 228701.
- [7] Petter Holme, Beom Jun Kim . Growing scale-free networks with tunable clustering[J]. Physical Review E, 2002, 65 (2) : 026107.
- [8] Joern Davidsen, Holger Ebel, et al. Emergence of a small world from local interactions: Modeling acquaintance networks[J]. Physical Review Letters, 2002, 88 (12) : 128701.
- [9] Mark Newman, Steven Henry Strogatz et al. Random graphs with arbitrary degree distributions and their applications[J]. Physical Review E , 2001, 64 (2) : 026118.
- [10] Jon Kleinberg, Steve Lawrence . The structure of the web[J]. Science, 2001, 294: 1849.
- [11] Filippo Menczer . Evolution of document networks[J]. Proceedings of the National Academy of Sciences, 2004, 101 (suppl 1) : 5261-5265.
- [12] Trevor Fenner, Mark Levene et al. A stochastic model for the evolution of the web allowing link deletion[J]. ACM Transactions on Internet Technology, 2006, 6 (2) : 117-130.
- [13] Xue-Qi Cheng, Fu-Xin Ren, et al. Triangular clustering in document networks[J]. New Journal of Physics, 2009, 11 (3) : 033019.
- [14] Fu-Xin Ren, Hua-Wei Shen et al. Modeling the clustering in citation networks[J]. Physica A : Statistical Mechanics and its Applications, 2012, 391 (12) : 3533-3539.
- [15] Lars Backstrom, Dan Huttenlocher et al. Group formation in large social networks: membership, growth, and evolution. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.
- [16] James Coleman . Foundations of Social Theory. Harvard, 1990.
- [17] Mark Granovetter. The strength of weak ties: A network theory revisited[J]. Sociological theory, 1983, 1 (1) : 201-233.
- [18] Johan Ugander, Lars Backstrom, et al. Structural diversity in social contagion[J]. Proceedings of the National Academy of Sciences, 2012, 109 (16) : 5962-5966.
- [19] Jure Leskovec, Daniel Huttenlocher, et al. Predicting positive and negative links in online social networks. Proceedings of the 19th international conference on World wide web. ACM, 2010.
- [20] Daniel Gruhl, Ramanathan Guha, et al. Information diffusion through blogspace. Proceedings of the 13th international conference on World Wide Web. ACM, 2004.

- [21] Fritz Heider. Attitudes and cognitive organization[J]. Journal of Psychology, 1946, 21: 107-112.
- [22] Dorwin Cartwright, Frank Harary . Structure balance: A generalization of Heider's theory[J]. Psychological Review , 1956, 63: 277-293.
- [23] Bo Yang, Xindong You, et al. Progress on analysis for detecting evolutionary community structure in complex dynamical networks[J]. Application Research of Computers, 2013, 30 (5) : 1292-1296.
- [24] John Hopcroft, Omar Khan, et al. Tracking evolving communities in large linked networks. Proceedings of the national academy of sciences of the United States of America[J], 2004, 101 (Suppl 1) : 5249-5253.
- [25] Derek Greene, Dónal Doyle, et al. Tracking the evolution of communities in dynamic social networks. Proceedings of the 2nd International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, 2010: 176-183.
- [26] Mansoureh Takaffoli, Farzad Sangi, et al. Community evolution mining in dynamic social networks[J]. Procedia Social and Behavioral Sciences, 2011, 22 (10) : 49-58.
- [27] Gergely Palla, Albert-László Barabási, et al. Quantifying social group evolution[J]. Nature, 2007, 446 (7136) : 664-667.
- [28] Deepayan Chakrabarti, Ravi Kumar, et al. Evolutionary clustering. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.
- [29] Yun Chi, Xiaodan Song et al. Evolutionary spectral clustering by incorporating temporal smoothness. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Dis-covery and Data Mining, 2007.
- [30] Lei Tang, Huan Liu et al. Identifying involving groups in dynamic multimode networks[J]. IEEE Transation on Knowledge and Data Engineering, 2012, 24 (1) : 72-85.
- [31] Yu-Ru Lin, Yun Chi, et al. FacetNet: A framework for analyzing communities and their evolutions in dynamic networks. Proceedings of the 17th International Conference on World Wide Web, ACM, 2008.
- [32] Teresa M. Selee, Tamara G. Kolda, et al. Extracting clusters from large datasets with multiple similarity measures using IMSCAND[J]. CSRI Summer Proceedings, 2007: 87-103.
- [33] Renaud Lambiotte, Jean-Charles Delvenne, et al. Laplacian dynamics and multiscale modular structure in networks. arXiv preprint arXiv: 0812.1770.
- [34] Peter J. Mucha, Thomas Richardson et al. Community structure in time-dependent, multiscale, and multiplex networks[J]. Science, 2010, 328 (5980) : 876-878.

Analysis of User Behavior

5.1 Introduction

As an emergent information technology, online social network has become an important part in people's daily life. By assessing different types of social network technologies, services and applications, people select the suitable one to achieve such goals as social intercourse, entertainment and information access. Social network user behavior refers to users' adoption and usage of social network services based on the comprehensive assessment of their needs, social influences and technological characteristics of social network.

User behavior is an important part in the research on online social network. Existing researches are mainly carried out based on the following two thoughts. The first group considers the online social network as a specific information technology and researches the adoption behavior, refusal behavior and user loyalty for online social network technology. The second group considers the social network as a platform that provides a wide selection of services and applications and researches the characteristics and laws presented in the usage of the services and applications.

Considering online social network as a specific information technology, researchers classify social network user behavior into several progressive levels such as adoption, usage and loyalty, and consider that adoption is the antecedent of actual user usage behavior, while loyalty is the user's trust in, dependence on and commitment to social network technology and the continuous usage behavior arising therefrom ^[2,8,13,19]. Based on

the above thoughts, researchers employ such classical behavioral research theories as Technology Acceptance Model, Theory of Planned Behavior, Theory of Expectation Confirmation and Flow Theory to explore the influence of such factors as demographic variables, personality traits, emotional factors, cognitive factors, motivation factors, social environment, physical environment and technological environment on the adoption and loyalty of online social network users.

Considering online social network as a platform that provides various services and applications, researchers have conducted extensive researches of individual usage behavior including self-presentation, microblog posting, search, browse and comments, as well as social interaction behavior including relationship establishment and content selection^[26,29,33,45,47]. They employ statistical methods, econometrics methods, queuing theory, etc. to analyze the distribution of social network user behavior and its temporal and spatial laws, thereby revealing the inherent mechanism of the content generation behavior and content consumption behavior of online social network, and extensively exploring the relationship selection laws, content selection laws and temporal laws of social interaction in the online social network.

This chapter is organized as follows. Introduce the classical theories, principal methods, fundamental process and research conclusions on social network user usage behavior with adoption, usage and loyalty as the main clues; introduce the influential factors, modeling methods and verification process for adoption behavior and user loyalty of social network; introduce general laws of social network user usage behavior and the modeling methods for social network content generation behavior and content consumption behavior; introduce the analysis methods for relationship selection behavior, content selection behavior and the temporal laws of social interaction. Finally, a summary is given for this chapter.

5.2 Online Social Network User Adoption and Loyalty

5.2.1 Online Social Network User Adoption

Online social network adoption refers to users' adoption behavior of online social network services based on the comprehensive assessment of their needs and motivations, social influences and technological characteristics of online social network. According to the diffusion theory of innovations (DTI) provided by Rogers, it is essential for subsequent

diffusion of online social network to be adopted and tried by as many users as possible in the early stage. Currently, researchers have used multiple theories to reveal the adoption behavior mechanism of online social network users, in which Technology Acceptance Model (TAM) and Theory of Planned Behavior (TPB) are the most popular ones.

1. Online Social Network User Adoption Models Based on the TAM

TAM, proposed by Davis Fred^[1], is one of the most classic models in current research field of information system. As shown in Figure 5-1, TAM assumed that users' actual adoption and usage behavior are directly and positively influenced by their use intention, which is influenced by their use attitude and Perceived Usefulness (PU); use attitude is influenced by PU and Perceived Ease of Use (PEOU), while PEOU influences PU to a certain degree. In TAM, PU refers to the degree to which the user believes the technology will increase his or her work performance and PEOU refers to the degree to which a user believes that a specific information system is effortless to use. TAM, after its establishment, is widely used to explain and predict the user adoption and acceptance behavior of new technologies, products and services, and obtains good results therein. Therefore, TAM is applicable for researching adoption behavior in online social network.

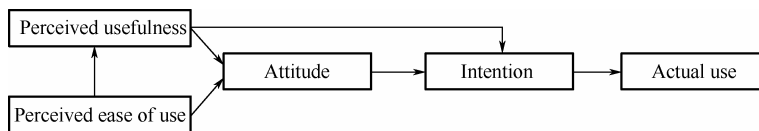


Figure 5-1 TAM proposed by Davis

When researching adoption behavior in online social network based on TAM, it is necessary to extend the traditional TAM as it only considers influences from four internal psychological factors but neglects influences from other internal psychological factors such as users' emotion and personality traits as well as external social factors such as social regulations and interpersonal influence. For example, Kwon Ohbyung et al.^[2] researched user adoption behavior in online social network based on TAM with basic process as follows.

Step 1: Reference review and model construction. Based on review and analysis on relevant reference systems of TAM, social identity theory, altruism, telepresence, etc., a research model is proposed as shown in Figure 5-2 and the following hypotheses on relations between various variables were inferred and demonstrated: Social identity (SI)

will have a positive effect on PEOU / PU / PE (perceived encouragement) of a social network service (H1a, H1b, H1c); altruism (ALT) will have a positive effect on PEOU / PU / PE of a social network service (H2a, H2b, H2c); telepresence will have a positive effect on PEOU / PU / PE of a social network service (H3a, H3b, H3c); PEOU will have a positive effect on PU of a social network service (H4); PE will have a positive effect on PU of a social network service (H5); PEOU will have a positive effect on actual use (AU) of a social network service (H6); PU will have a positive effect on AU of a social network service (H7); and PE will have a positive effect on AU of a social network service. In this research model, SI is defined as the degree that an individual's knowledge of belonging to a certain social group; ALT is defined as the inclination of helping others; telepresence (TELE) means an individual's feeling of being present in a virtual environment generated by media; and PE refers to the encouragement and support from others perceived by an individual.

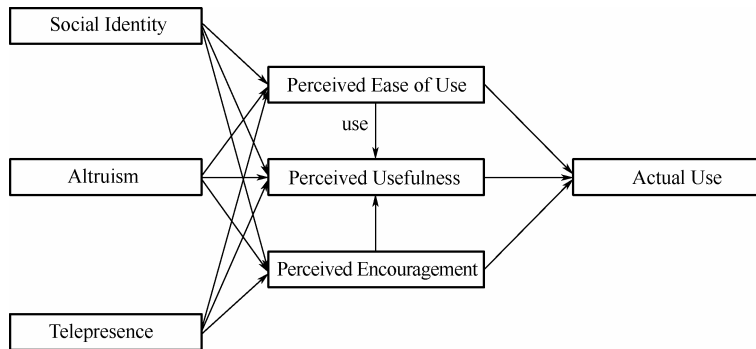


Figure 5-2 Research model of Kwon Ohbyung et al. (2010)

Step 2: Scale development and data acquisition. Measurement items for PEOU, PU, PE, SI, ALT, TELE, AU scales and other variables were developed on the basis of related research achievements (see Table 5-1), which is used to conduct a survey on SNS user group in employees of South Korea companies and collect 229 pieces of sample data available for verifying model assumption. These samples consist of 53.4% male and 66.2% of samples were at the age of 20 ~ 29.

Table 5-1 Measurement items used by Kwon Ohbyung et al. (2010)

Variable	Measurement items	Scale
PU	Using the SNS enables me acquire more information or meet more people	1-2-3-4-5-6-7
	Using the SNS would improve my efficiency in sharing information and connecting with others others	1-2-3-4-5-6-7
	The SNS is a useful service for communication	1-2-3-4-5-6-7

(To be continued)

Continued table

Variable	Measurement items	Scale
	The SNS is a useful service for interaction of members	1-2-3-4-5-6-7
PEOU	Learning to use the SNS is easy for me	1-2-3-4-5-6-7
	The process of using the SNS is clear and understandable	1-2-3-4-5-6-7
	I find it easy to use the SNS	1-2-3-4-5-6-7
PE	People whom I meet in the SNS tend to give me affirmative evaluation	1-2-3-4-5-6-7
	People whom I meet in the SNS tend to be satisfied with me	1-2-3-4-5-6-7
	People whom I meet in the SNS give me great encouragement	1-2-3-4-5-6-7
	People whom I meet in the SNS tend to be aware of my existence	1-2-3-4-5-6-7
SI	As a member of the community, my position is very important to me	1-2-3-4-5-6-7
	As a member of the community, I am the type of person who likes to engage in my community	1-2-3-4-5-6-7
	Activities in my community are an important part in my life	1-2-3-4-5-6-7
ALT	I tend to encourage people who are in a real crisis or need	1-2-3-4-5-6-7
	I usually help people who ask me for solution	1-2-3-4-5-6-7
	I give congratulation when people tell me good news	1-2-3-4-5-6-7
TELE	When exiting the SNS, I felt like I actually met other people	1-2-3-4-5-6-7
	I felt that the SNS creates a new world	1-2-3-4-5-6-7
	While using with the SNS, I felt I was in a different society	1-2-3-4-5-6-7
	While using with the SNS, the SNS world was more real or present to me compared to the "real world"	1-2-3-4-5-6-7
AU	I tend to use the SNS frequently	1-2-3-4-5-6-7
	I spend a lot of time on SNS	1-2-3-4-5-6-7
	I exerted myself to SNS	1-2-3-4-5-6-7

Note: Measurement of all variables adopts seven-level Likert scale (1-"Strongly disagree", 7-"Strongly agree").

Step 3: Data analysis. The research mainly used LISREL 8.7 and other softwares to conduct scale reliability and validity analysis and structural equation modeling (SEM) approach. Reliability and validity analysis means the process of verifying reliability and validity through confirmatory factor analysis (CFA), calculating Cronbach's α and other methods. As shown in factor analysis, the scale has good validity as factor loads of most measurement items are larger than 0.7. As shown in Table 5-2, the scale has good reliability as all Cronbach's α are larger than 0.7. Structural equation modeling (SEM) means the process of verifying complex model with multiple variables through Partial Least Squares (PLS) and other parameter evaluation methods to discover the significant relations between variables. Analysis results are as follows:

(1) Except that AGFI is slightly smaller than reference value, all remaining index values reflecting the model fitness are larger than the reference value, indicating good fitness of the research model (see Table 5-3).

(2) Except that research hypothesis H1a, H2b and H3b are untenable, the relation hypotheses between all remaining variables are tenable (see Table 5-4).

Table 5-2 Cronbach's α of variables used by Kwon Ohbyung et al. (2010)

Variables	SI	ALT	TELE	PU	PEOU	PE	AU
α	0.81	0.88	0.80	0.90	0.86	0.89	0.78

Table 5-3 Fitting degree for model used by Kwon Ohbyung et al. (2010)

Index of fitting degree	χ^2 / df	GFI	AGFI	NFI	NNFI	CFI	RMSEA
Index value	2.25	0.84	0.79	0.93	0.95	0.96	0.07
Reference value	≤ 3	≥ 0.9	≥ 0.8	≥ 0.9	≥ 0.9	≥ 0.9	≤ 0.10

Note: See *Structural Equation Modeling – Operation and Application of AMOS* written by Minglong Wu for computing formula of various indexes.

Table 5-4 Estimated value for model parameter and verification results for hypothesis used by Kwon Ohbyung et al. (2010)

Dependent variables	Independent variables	Standardized coefficients (β)	Supported or not
PEOU ($R^2=0.27$)	SI	0.12	H1a (No)
	ALT	0.22**	H2a (Yes)
	TELE	0.39**	H3a (Yes)
PE ($R^2=0.50$)	SI	0.24**	H1c (Yes)
	ALT	0.29**	H2c (Yes)
	TELE	0.46**	H3c (Yes)
PU ($R^2=0.48$)	PEOU	0.44**	H4 (Yes)
	PE	0.47**	H5 (Yes)
	SI	0.01**	H1b (Yes)
	ALT	-0.11	H2b (No)
	TELE	0.01	H3b (No)
AU ($R^2=0.60$)	PEOU	0.39**	H6 (Yes)
	PE	0.31*	H8 (Yes)
	PU	0.21**	H7 (Yes)
Note: * $p < 0.05$, ** $p < 0.01$			

Step 4: Conclusion and Discussion. According to data analysis results, TELE and ALT only influence PU indirectly through PEOU and PE, while SI has significant direct effect on PEOU and PE. In addition, this research sees SNS as a kind of relationship-oriented information system and SNS may become a kind of task-oriented information system. According to users and their orientation for SNS, information systems can be classified into the following four types as shown in Figure 5-3.

Orientation	Relationship-oriented	Relationship-oriented tools	Collective emotion tools
	Task-oriented	Legacy information sharing tools	Collective intelligence tools
		Individual	Collective
		Users	

Figure 5-3 Information system classifications based on orientation and users

In general, researchers conducted the empirical research on adoption behavior in online social network based on TAM by following the preceding steps, and the main differences among this research are new variables added when extending TAM. For example, Ernst Claus-Peter et al. ^[3] constructed a SNS user adoption model by extending the TAM with two variables: perceived enjoyment and perceived belonging (see Figure 5-4). They verified relation hypothesis between all variables in the model by empirical research. Nikou Shahrokh ^[4] proposed a mobile SNS user adoption model by integrating mobility, critical mass, use context, social influence, habit and other variables into TAM (see Figure 5-5). As shown in empirical research, mobility influences use intention indirectly through PEOU; critical mass influences use intention indirectly through social influence; and social influence, PEOU and habit have significant direct effect on use intention. In addition, Deb Sledgianowski et al. ^[5] proposed a research model based on TAM by adding four variables including perceived playfulness, perceived normative pressure, trust and critical mass as the direct influence factors of adoption intention. According to empirical research, they discovered that these four factors all have significant influences on adoption intention, and perceived playfulness has significant direct influence on actual adoption. Bao Dai et al. ^[6] constructed a SNS user adoption model by adding perceived popularity into TAM to reflect social influence, and used perceived popularity as the reason for PU, PEOU and use attitude. The empirical research verified relation hypothesis among all variables in the model.

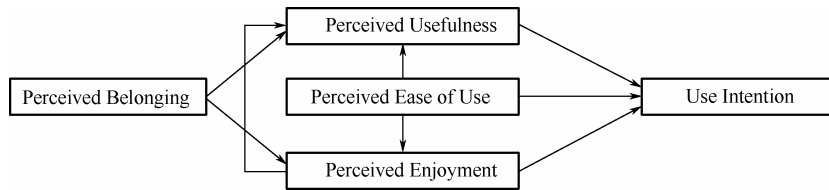


Figure 5-4 Research model used by Ernst Claus-Peter et al. (2013)

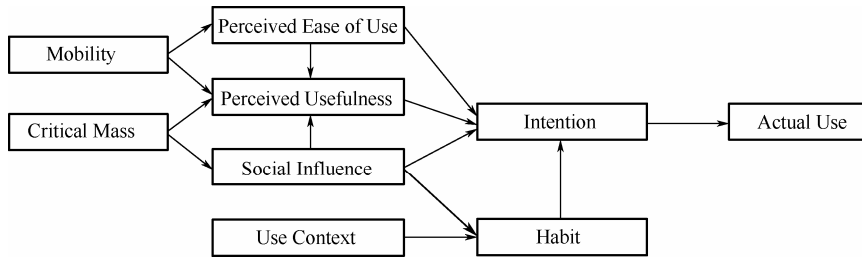


Figure 5-5 Mobile SNS user adoption model

2. Online Social Network User Adoption Models Based on TPB

TPB proposed by Icek Ajzen^[7] is a classical theory applied extensively to the research on human behaviors. As shown in Figure 5-6, the TPB posits that individuals' behaviors are directly driven by behavioral intention which is determined by attitude, perceived behavioral control (PBC) and subjective norm (SN). Attitude is defined as individuals' positive or negative evaluations about performing a target behavior. PBC refers to an individual's perception of his/her resources and capacity to perform a target behavior. Subjective norm reflects an individual's perceptions of social pressure from important referents to perform or not to perform the behavior.

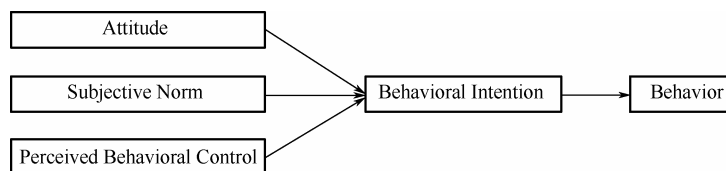


Figure 5-6 TPB proposed by Icek Ajzen

Considering that TPB shows the effects of both individuals' internal psychological factor (attitude and PBC) and external social influence factors (subjective norm) on behavioral intention, it may be suitable for explaining user adoption behavior for online social network. Overall, researchers usually apply TPB to research on user adoption

behavior in online social network in the following steps. First, integrate some new variables into model shown in Figure 5-6 to construct new research model; second, develop scale to conduct investigation and obtain empirical research data; third, test the model by using statistical methods, such as SEM approach and regression analysis, and finally draw conclusions according to analysis results and make discussion. For example, Baker Rosland et al. [8] researched user adoption behavior in online social network on the basis of TPB, whose research process is as follows.

Step 1: Reference review and model construction. Based on the review of TPB, social identity theory, self-categorization theory and research on self-esteem in personality psychology, they pointed out that factors influencing SNS user adoption behavior include attitude, PBC, SN, group norm and self-esteem, and thereby constructed the regression model with the above factors as independent variables and adoption intention as dependent variables. Group norm refers to behavioral principles observed by the membership of a group. Group norm has influence on behaviors of group members as an individual has to observe group norm to gain recognition from the group. Self-esteem, as an important part of self-concept, refers to an overall positive or negative evaluation of the self, and has fundamental influence on individual behavior.

Step 2: Scale development and data acquisition. They developed a scale for measuring various variables in the model on the basis of previous related research, and then conducted investigation on Australian juniors and obtained 160 pieces of valid sample data (schoolboy accounts for 36% and has an average age of 14.36). Attitude scale was a 7-level semantic differential response scale with five items (e.g., unpleasant 1 - 2 - 3 - 4 - 5 - 6 - 7 pleasant). Subjective norm scale was a 7-level Likert scale with two items (e.g., “Most people who are important to me want me to socialize online in the future by SNS like Facebook”; strongly disagree 1 - 2 - 3 - 4 - 5 - 6 - 7 strongly agree). PBC scale was a 7-level Likert scale with four items (e.g., “I have complete control over whether to socialize online in the future by SNS like Facebook”; strongly disagree 1 - 2 - 3 - 4 - 5 - 6 - 7 strongly agree). Group norm scale was a 7-level Likert scale with four items (e.g., “Most of my friends will socialize online in the future by SNS like Facebook”; strongly disagree 1 - 2 - 3 - 4 - 5 - 6 - 7 strongly agree). Self-esteem scale proposed by Rosenberg included 1- items (e.g., “I think that I have a number of good qualities”; strongly disagree 1 - 2 - 3 - 4 strongly agree).

Step 3: Data analysis results. Conduct descriptive statistical analysis, correlation analysis and hierarchical regression analysis by SPSS (see Table 5-5 and Table 5-6 for results). Descriptive statistical analysis gives mean (M) and standard deviation (SD);

correlation analysis gives correlation coefficient between variables and the significant level; and hierarchical regression analysis reveals the influence of various factors as independent variables on use intention for SNS as dependent variable. As shown in Table 5-5, attitude, PBC, SN and group norm have significant positive correlation with use intention for SNS. As further shown in Table 5-6, attitude, PBC and group norm have significant influence on use intention for SNS, and their influence degrees from big to small are group norm, attitude and PBC.

**Table 5-5 Mean, SD, and correlation coefficient
for variables used by Baker & Rosland et al. (2010)**

Variable	M	SD	1	2	3	4	5	6
Attitude	4.73	1.28	1	0.47***	0.26**	0.51***	0.43***	0.00
Subjective norm	4.28	1.42		1	0.25**	0.47***	0.61***	-.05
PBC	5.35	1.32			1	0.29***	0.24**	-.18*
Intention	3.53	1.69				1	0.59***	0.11
Group norm	4.58	1.33					1	0.09
Self-esteem	2.12	0.54						1

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5-6 Results of hierarchical regression analysis provided by Baker & Rosland et al. (2010)

Variable	Unstandardized regression coefficients B	Standardized Error (SE)	Standardized regression coefficients β	R^2
Step 1				0.35***
Attitude	0.45	0.10	0.34***	
Subjective norm	0.33	0.09	0.28***	
PBC	0.19	0.09	0.14*	
Step 2				
Attitude	0.36	0.09	0.27***	0.45***
Subjective norm	0.10	0.10	0.08	
PBC	0.17	0.08	0.13*	
Group norm	0.49	0.10	0.38***	
Self-esteem	0.32	0.19	0.10	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Step 4: Conclusion and Discussion. According to statistical analysis results, attitude and PBC have significant influence on teenagers' use intention for SNS, thus this research supports the validity of behavior prediction by TPB to a certain degree. However, SN has non-significant

influence on teenagers' use intention for SNS, which is not in conformity with many researches conducted based on TPB. The reason may be because, compared with SN, the newly-added variable, i.e. group norm, can better reflect the influence of pressure from social norm on adoption behavior. Such inference can be supported by the phenomenon that the influence of SN on use intention turns from significant to non-significant after group norm is added into regression model. In addition, self-esteem has non-significant direct influence on use intention for SNS, which is also not in conformity with previous relevant researches. The reason may be that self-esteem needs other intermediary variables to indirectly influence behavior intention, which needs further verification by research in the future.

Except for the research of Baker Rosland et al. (2010), there are some other empirical researches on user adoption behavior in online social network based on TPB. They have similar procedures with that of research by Baker Rosland et al. (2010) with the main differences in added variables and statistical process methods. For instance, Emma Pelling et al.^[9] constructed a prediction model for use intention for SNS users by adding self-identity, belongingness, age, past use and other new variables to TPB. As shown in results of hierarchical regression analysis method, attitude, SN and self-identity have significant positive influence on high-level use intention for SNS. Yaping Chang et al.^[10] constructed a research model by using 5 kinds of adoption motive as the antecedent for attitude in their research on adoption behavior of SNS users in China based on TPB. As shown in results of SEM analysis, information, entertainment, new acquaintance and conformity significantly influence the use attitude for SNS, and users' attitude, PBC and SN significantly influence adoption intention of SNS user. Goh Say Leng et al.^[11] proposed an adoption model of SNS user by integrating TPB and TAM (see Figure 5-7). As shown in results of SEM analysis, PU, attitude and PBC have significant positive influence on adoption intention, and adoption intention significantly influence actual adoption.

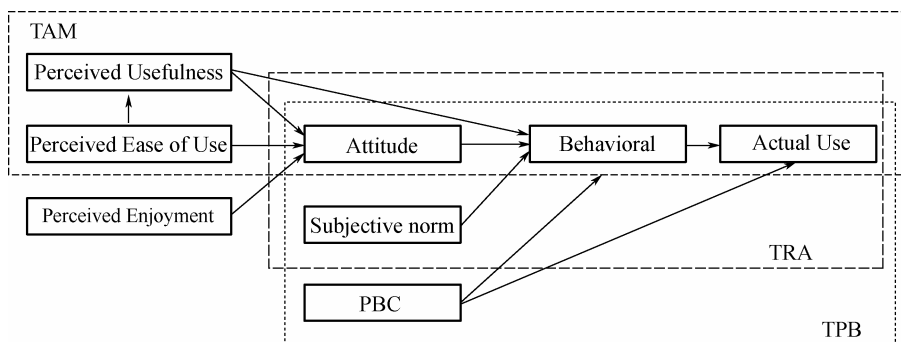


Figure 5-7 Research model used by Goh Say Leng et al. (2011)

5.2.2 Online Social Network User Loyalty

User retention is of great importance to online social network service providers because the eventual success of online social network depends on continuance usage rather than on first-time use. However, competition pressure brought by constant emergence of various new network services makes it more difficult to maintain the online social network user loyalty. Therefore, it is very important to research the mechanism of online social network user loyalty. So far, multiple kinds of theories have been applied to user loyalty research, including the widely-accepted expectation confirmation theory (ECT) and the flow theory.

1. Online Social Network User Loyalty Models Based on ECT

ECT first proposed by Oliver (1980) is the basic theory for researching consumer satisfaction degree. As stated in ECT, consumers have certain expectations for the products or services to be purchased, and compare perceived actual performance after purchase with previous expectations to evaluate the confirmation. If yes, they will be satisfied; otherwise, they will be unsatisfied. Consumer satisfaction degree will further influence the intention and behavior of repurchasing such products or services. Anol Bhattacharjee ^[12], by combining ECT and characteristics of information system, proposed expectation confirmation model of information system continuance (ECM-ISC). As shown in Figure 5-8, ECM-ISC posits the expectation confirmation of information system users influences their intention to continue using the information system by influencing their PU and satisfaction degree on it. Currently, ECM-ISC is widely used in research of user loyalty or continuous use behavior and become one of the most famous theories in the information technology research field.

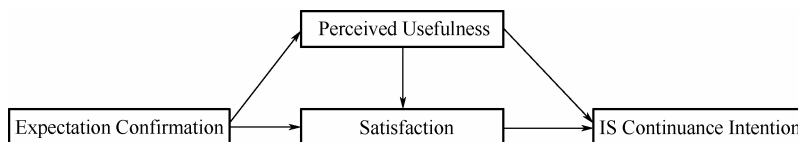


Figure 5-8 Expectation confirmation model used by Anol Bhattacharjee (2001)

For research on user loyalty in online social network based on ECT, a usual method is to add other variables to ECM-ISC, verify such model according to the process of scale

development → data acquisition → data process, and finally reach a conclusion according to analysis results for discussion. Young Sik Kang et al. ^[13] researched the user loyalty in online social network based on ECT according to the continuous usage behavior in Cyworld, the biggest social network in South Korea, and the research process and conclusion are as follows.

Step 1: Reference review and model hypothesis. After reviewing the reference about ECT and combining research results on regret theory and self-image congruity, Young Sik Kang et al. proposed the research model as shown in Figure 5-9, and conducted detailed inference and demonstration on relation hypotheses among various variables in the model, including the additional variables of regret, self-image congruity, perceived enjoyment and past use. Regret refers to the degree that users regret their selection of using Cyworld; self-image congruity refers to the degree that users deem using Cyworld as congruent with their images; perceived enjoyment refers to the degree that users deem using Cyworld as an enjoyment experience; past use refers to the situation that users use Cyworld in the past period. The model includes the following 13 research hypotheses: Satisfaction positively influences continuous usage intention of online social network service (H1); Confirmation of expectations positively influences satisfaction with online social network service (H2); PU positively influences satisfaction with online social network service (H3); Perceived enjoyment positively influences satisfaction with online social network service (H4); PU positively influences continuous usage intention of online social network service (H5); Perceived enjoyment positively influences continuous usage intention of online social network service (H6); Confirmation of expectations positively influences PU (H7); Confirmation of expectations positively influences perceived enjoyment (H8); Self-image congruity with online social network service positively influences continuous usage intention of online social network service (H9); Self-image congruity with online social network service positively influences PU (H10); Self-image congruity with online social network service positively influences perceived enjoyment (H11); Regret negatively influences satisfaction with online social network service use (H12); Regret negatively influences continuous usage intention of online social network service (H13).

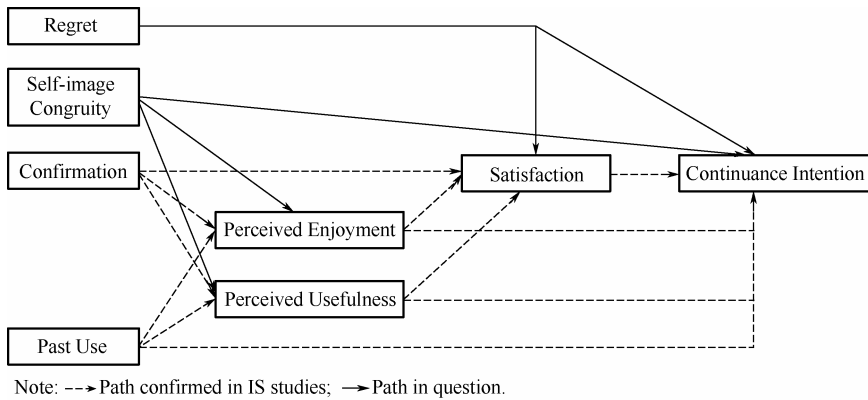


Figure 5-9 Research model used by Young Sik Kang et al. (2009)

Step 2: Scale development and data acquisition. Based on previous research results, Young Sik Kang et al. developed the scale for measuring various variables in the above model (see Table 5-7), thereby conducted a field survey on Cyworld users in university and obtained 349 pieces of valid sample data (47.9% from male, 72.8% from users between their 20 ~ 29 and 81.8% from users having over 1 year of use experience).

Table 5-7 Survey measurement items used by Young Sik Kang et al. (2009)

Variables	Items	Scale
Regret	I regret the selection of using Cyworld	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	I very much regret the selection of using Cyworld	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	I should have selected other social networks	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Self-image congruity	Visiting Cyworld helps maintain my image and character	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	Visiting Cyworld helps in reflecting who I am	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	Visiting Cyworld fits well with my image	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Confirmation	My experience with using Cyworld was better than I expected	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	The service level provided by Cyworld was better than I expected	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	Overall, most of my expectations from using Cyworld were confirmed	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Perceived enjoyment	Using Cyworld is enjoyable	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	Using Cyworld is pleasurable	Strongly disagree 1-2-3-4-5-6-7 Strongly agree

(To be continued)

Continued table

Variables	Items	Scale
	I think it is interesting to use Cyworld	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
PU	Using Cyworld improves my productivity in managing personal information	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	Using Cyworld improves my efficiency in managing my personal information	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	Overall, Cyworld is useful in managing my personal information	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Satisfaction	How do you feel about your overall experience of using Cyworld:	Very dissatisfied 1-2-3-4-5-6-7 Very satisfied
		Very displeased 1-2-3-4-5-6-7 Very pleased
		Very frustrated 1-2-3-4-5-6-7 Very contented
		Very terrible 1-2-3-4-5-6-7 Very delighted
Continuous usage intention	I intend to continue using Cyworld rather than discontinue its use	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	My intentions are to continue using Cyworld rather than use any alternative means	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	If possible, I would like to discontinue my use of Cyworld (R)	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Past use	On average, how frequently have you visited Cyworld over the past month? 1 = less than once a month; 2 = once a month; 3 = a few times a month; 4 = a few times a week; 5 = about once a day; and 6 = several times a day	
	On average, how much time have you spent per day visiting Cyworld over the past month? 1. Less than 10 min; 2. 10 ~ 20 min; 3. 20 ~ 30 min; 4. 30 min ~ 1 h; 5. 1 ~ 2 h; 6. 2 ~ 3 h; 7. More than 3 h	

Step 3: Data analysis results. Verify the reliability and validity degree of the scale by CFA and model hypothesis by SEM. The results of descriptive statistical analysis and CFA are as shown in Table 5-8. Convergent validity is acceptable as the composite reliability (CR) of all variables is higher than 0.70 and the average variance extracted (AVE) is higher than 0.50. In addition, the square roots of all AVEs are larger than all other cross correlations, suggesting adequate discriminant validity. As shown in Table 5-9, all research hypotheses are verified except for H2 and H12.

Table 5-8 Variable mean, SD, CR and AVE used by Young Sik Kang et al. (2009)

Variable	M	SD	CR	1	2	3	4	5	6	7	8
Regret	2.61	1.20	0.92	0.89							
Self-image congruity	4.29	1.45	0.93	-0.29	0.91						
Confirmation	5.00	1.22	0.94	-0.46	0.45	0.91					
Perceived enjoyment	4.70	1.29	0.95	-0.33	0.56	0.58	0.93				

(To be continued)

Continued table

Variable	M	SD	CR	1	2	3	4	5	6	7	8
PU	5.16	1.16	0.94	-0.35	0.42	0.52	0.49	0.90			
Satisfaction	4.52	1.32	0.96	-0.20	0.30	0.34	0.43	0.33	0.94		
Continuous usage intention	5.11	1.38	0.91	-0.55	0.46	0.46	0.53	0.54	0.33	0.87	
Past use	3.69	1.68	0.87	-0.23	0.38	0.38	0.43	0.25	0.18	0.44	0.88

Note: Values on the diagonal are the square roots of AVE.

Table 5-9 Estimated value for model parameter and hypothesis verification results used by Young Sik Kang et al. (2009)

Effects	Causes	Estimated value for model parameter	Hypothesis verification results (Supported or not?)
PU	Past use	0.018	(No)
	Confirmation	0.411 ^{***}	H7 (Yes)
	Self-image congruity	0.231 ^{***}	H10 (Yes)
Perceived enjoyment	Past use	0.177 ^{***}	
	Confirmation	0.365 ^{***}	H8 (Yes)
	Self-image congruity	0.331 ^{***}	H11 (Yes)
Satisfaction	Confirmation	0.098	H2 (No)
	PU	0.131 ^{**}	H3 (Yes)
	Perceived enjoyment	0.305 ^{***}	H4 (Yes)
	Regret	-0.005	H12 (No)
Continuous usage intention	Past use	0.203 ^{***}	
	PU	0.243 ^{***}	H5 (Yes)
	Perceived enjoyment	0.140 ^{**}	H6 (Yes)
	Satisfaction	0.063 [*]	H1 (Yes)
	Self-image congruity	0.080 [*]	H9 (Yes)
	Regret	-0.338 ^{***}	H13 (Yes)

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Step 4: Conclusion and Discussion. Overall, this research contributes to the online social network service continuance reference by demonstrating that self-image congruity and regret can be seamlessly incorporated into the ECM perspective to explain continuous usage behavior of online social network service. Opposite to expectation, the negative effect of regret on satisfaction is not significant, which could be partly because the impact of regret on satisfaction is found to be significant only under negative confirmation (Taylor, 1997), but most respondents' expectations from using Cyworld in this sample are confirmed as positive ($M=5.00$, $SD=1.22$). In addition, the insignificant effect of confirmation on

satisfaction and the weak effect of satisfaction on continuous usage intention are possibly due to effect from behavioral habit [Kim et al. (2005), Limayem et al (2007)]. In this research, about 80% of the respondents visited Cyworld a few times a week and about 75% used it more than 10 min per day over the last month, indicating the habitual use of many respondents.

Apart from the empirical research of Young Sik Kang et al. (2009) on the continuous usage behavior of online social network based on ECT, there are many other similar research results. Those results have basically the same research process with the research of Young Sik Kang et al. (2009), but with different variables when constructing research model. For instance, Guopeng Yin et al. ^[14] developed a research model by adding perceived enjoyment, structural embeddedness, perceived privacy risk, and past use into the ECM (see Figure 5-10). The results show that PU, structural embeddedness and satisfaction have significant positive effect on continuous usage intention. Yao Chen et al. ^[15] constructed their research model by integrating PEOU, perceived playfulness and perceived switching cost into the ECM (see Figure 5-11). The results show that PEOU, PU, perceived playfulness and perceived switching cost directly influence continuous usage intention in SNS. In addition, Shin Soo et al. ^[16] proposed an amended model of ECM by replacing PU with perceived characteristics of innovation, which is derived from the diffusion theory of innovations (DTI). The results show that perceived characteristics of innovation is a significant influencing factor on continuous usage intention for online social network . Qian Li ^[17] proposed a research model by integrating the information system success model and the ECM. The results show that system quality, information quality and perceived service accessibility significantly influence satisfaction and then influence continuous usage intention for SNS.

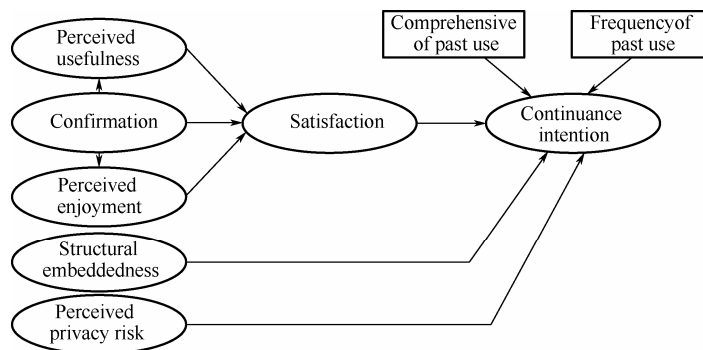


Figure 5-10 Research model used by Guopeng Yin (2010)

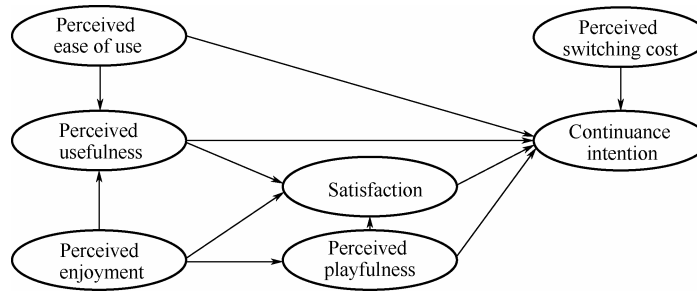


Figure 5-11 Research model used by Yao Chen (2011)

2. Online Social Network Users' Loyalty Models Based on the Flow Theory

Flow theory proposed by Mihaly Csikszentmihalyi et al.^[18] is an important theory in user experience research. Flow experience refers to “the holistic experience that people feel when they act with total involvement”, characterized by high concentration on the task at hand, a loss of self-consciousness, a distorted sense of time, internal enjoyment and so on. Flow experience is the optimal experience that may be obtained in many daily activities, as well as an autotelic experience, i.e. obtaining flow experience will become the purpose for the activity, thereby motivating participants.

Some researchers researched online social network users' loyalty based on the flow theory and proposed the research model, in which flow experience was taken as an antecedent of users' satisfaction or continuous usage intention. For instance, Zhou Tao et al.^[19] researched mobile SNS users' loyalty on the basis of the flow theory, and the process of this research is as follows.

Step 1: Reference review and model construction. At the beginning, the authors reviewed and analyzed the theoretical basis of the flow theory, trust and information system success model, and proposed the research model for mobile SNS users' loyalty (see Figure 5-12). In the model, information quality refers to the accuracy, comprehensiveness and timeliness of information provided by mobile SNS service operator; system quality refers to the reliability, response speed and ease of use of mobile SNS platform. Flow experience, as a second-order factor, contains three dimensions, i.e. perceived enjoyment, perceived control and attention focus. The model contains the following hypotheses: The information quality of mobile SNS significantly influences user trust (H1); the information quality of mobile SNS significantly influences flow experience (H2); the system quality of mobile SNS significantly influences user trust (H3); the system quality of mobile SNS significantly

influences flow experience (H4); user trust significantly influences flow experience (H5); user trust significantly influences his/her loyalty for mobile SNS (H6); flow experience significantly influences user loyalty for mobile SNS (H7).

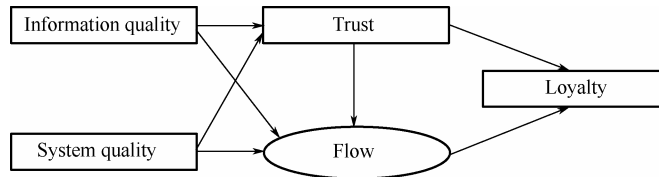


Figure 5-12 Research model used by Zhou Tao et al. (2010)

Step 2: Scale development and data acquisition. Based on previous research results, Zhou Tao et al. developed the scale for measuring various variables in the above model (see Table 5-10), thereby conducted a survey on students in a university in eastern China and obtained 305 pieces of valid sample data (57.4% from male and 83.6% from undergraduate students).

Table 5-10 Measurement items used by Zhou Tao et al. (2010)

Variable	Items	Scale
Information quality	The information provided by this mobile SNS is what I need.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	The information provided by this mobile SNS is accurate.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	The information provided by this mobile SNS is up-to-date.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	The information provided by this mobile SNS is comprehensive.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
System quality	This mobile SNS is reliable.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	This mobile SNS provides fast responses to my inquiries.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	This mobile SNS is easy to use.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	This mobile SNS provides good navigation functions.	Strongly disagree 1-2-3-4-5-6-7 strongly agree
Perceived enjoyment	I feel that using this mobile SNS is fun.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	I feel that using this mobile SNS is exciting.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	I feel that using this mobile SNS is enjoyable.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	I feel that using this mobile SNS is interesting.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Perceived control	When using this mobile SNS, I felt calm.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree

(To be continued)

Continued table

Variable	Items	Scale
Perceived control	When using this mobile SNS, I felt in control.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	When using this mobile SNS, I felt confused.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Attention focus	When using this mobile SNS, I was intensely absorbed in the activity.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	When using this mobile SNS, my attention was focused on the activity.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	When using this mobile SNS, I concentrated fully on the activity.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	When using this mobile SNS, I was deeply engrossed in the activity.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Trust	This mobile SNS has the necessary ability to fulfil its tasks.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	This mobile SNS will keep its promises.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	This mobile SNS is concerned with its users' interests.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
Loyalty	I will continue using this mobile SNS.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	I will recommend this mobile SNS to other users.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree
	When using mobile SNS, I consider this mobile SNS to be my first choice.	Strongly disagree 1-2-3-4-5-6-7 Strongly agree

Step 3: Data analysis. The research mainly used LISREL 8.7 and SPSS 13.0 to conduct CFA analysis to test the reliability and validity, and SEM analysis to test the model hypotheses. As shown in Table 5-11, all the AVE values exceed 0.5 and the CR values exceed 0.7, indicating good convergent validity of the scale; all Cronbach α values exceed 0.7, indicating good reliability of the scale. As shown in Table 5-12, all square roots of AVE are larger than its correlation coefficients with other variables, indicating good discriminant validity of the scale. As shown in Table 5-13, all model fitness indexes, except for GFI, are larger than reference value, indicating good fitness of the scale. As shown in Table 5-14, all estimated values for parameters reach significant level, indicating all 7 hypotheses are supported.

Table 5-11 AVE, CR and Cronbach α values used by Zhou Tao et al. (2010)

Variable	AVE	CR	α
Information quality	0.60	0.82	0.82
System quality	0.74	0.82	0.92
Perceived enjoyment	0.63	0.87	0.87
Perceived control	0.58	0.80	0.79
Attention focus	0.75	0.92	0.92
Trust	0.82	0.93	0.93
Loyalty	0.73	0.89	0.88

Table 5-12 Variable correlation coefficients and AVE used by Zhou Tao et al. (2010)

	1	2	3	4	5	6	7
Information quality	0.775						
System quality	0.443	0.857					
Perceived enjoyment enjoyment	0.631	0.429	0.793				
Perceived control	0.634	0.541	0.470	0.760			
Attention focus	0.377	0.466	0.279	0.686	0.864		
Trust	0.485	0.630	0.461	0.684	0.541	0.906	
Loyalty	0.426	0.436	0.485	0.652	0.578	0.616	0.853

Note: The square root of the AVE is shown on the diagonal.

Table 5-13 Model fitness used by Zhou Tao et al. (2010)

Fitness index	χ^2 / df	GFI	AGFI	NFI	NNFI	CFI	RMSEA
Actual values	2.62	0.897	0.853	0.961	0.966	0.973	0.068
Reference values	≤ 3	≥ 0.9	≥ 0.8	≥ 0.9	≥ 0.9	≥ 0.9	≤ 0.08

Table 5-14 Estimated value for model parameter and hypothesis verification results used by Zhou Tao et al. (2010)

Effects	Causes	Estimated value for model parameter	Hypothesis verification results (Supported or not?)
Trust	Information quality	0.26***	H1 (Yes)
	System quality	0.51***	H3 (Yes)
Flow experience	Information quality	0.38***	H2 (Yes)
	System quality	0.14*	H4 (Yes)
	Trust	0.45***	H5 (Yes)
Loyalty	Trust	0.20**	H6 (Yes)
	Flow experience	0.57***	H7 (Yes)

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Step 4: Conclusion and Discussion. According to the data analysis results, both information quality and system quality significantly influence users' trust and flow experience, while trust influences flow experience and loyalty and flow experience significantly influences user loyalty. Moreover, the results also show that perceived enjoyment, perceived control and attention focus all have high loadings on the second-order factor flow experience, indicating that it is appropriate to integrate the three dimensions into a reflective second-order factor. In addition, compared with information quality, system quality has a larger influence on user trust. Therefore, mobile SNS providers need to

attach importance to the reliability, response speed, navigability and ease of use of system performance and optimize human-machine interface design of mobile terminals, so as to improve user trust of SNS by overcoming the constraints of mobile terminals, such as small screen, low resolution and inconvenient input. On the other hand, compared with system quality, information quality has a larger influence on flow experience. Therefore, mobile service providers need to attach importance to the information quality and provide the latest, most accurate and most comprehensive information to users, so as to promote users' continuous usage behavior by improving user experience.

In addition to research of Zhou Tao (2010), there are some other researches on online social network users' loyalty based on the flow theory. For instance, Lin Hsu Chia ^[20] developed a research model to explain Facebook users' continuous usage by integrating the unified theory of adoption and use of technology (UTAUT), the expectation disconfirmation theory (EDT) and flow theory. Empirical verification on 482 pieces of valid sample data is conducted by SEM technology. Results show that flow experience not only indirectly influences continuous usage intention by satisfaction, but also has significantly direct influence on continuous usage intention. For another instance, Chang Ya Ping ^[21] researched continuous usage behavior of SNS users in China by constructing a model based on the ECM, social capital theory and flow theory. As shown in empirical research based on SEM, expectation conformation degree of SNS users has significantly positive influence on flow experience, which indirectly influences continuous usage intention by satisfaction, but the direct influence of flow experience on continuous usage intention for SNS failed the test. In addition, based on flow theory, Wu Yi et al. ^[22] and Chang Chiao Chen ^[23] also applied flow theory on the continuous usage behavior of social games and recreational application in SNS. Based on flow theory, Wu Yi et al. (2010) proposed that perceived enjoyment is the direct antecedent for user stickiness on recreational application in SNS, while perceived enjoyment is influenced by design factor (e.g., internal sociability and interactivity between applications) of recreational application in SNS. Empirical research supports the above relation hypothesis. Based on flow theory, Chang Chiao Chen (2013) constructed the research model with interactivity (human-machine interaction and social interaction), user value (tool value and entertainment value) and satisfaction as the antecedent variable of flow experience, and flow experience and satisfaction as the direct antecedent of social game. As shown in the empirical research results, relation hypothesis between various variables in the model is verified.

5.3 Individual Usage Behavior

5.3.1 General Usage Behavior

With the rapid development of forum, blog and microblogging and other online social networks, online social behavior of users shows more and more content and manifestation different from the past. Current research about user behavior in social network is mostly based on surveys or interviews. For example, Ryan Tracii et al. ^[24] investigated Facebook usage behavior of 1,324 users, and discovered that extrovert and nervous users tend to frequently use social network and spend more time on social network. Moore Kelly et al. ^[25] researched the general behavior rules of 219 college students in Facebook, and stated that the more experienced Facebook users were likely to spend more time on Facebook, post more photos, but disclose less personal information.

Along with the progress of social network applications and related technologies of data acquisition, it is possible to study online social network behavior based on large-scale user behavior data. For instance, based on the clickstream data in Orkut, Myspace, Hi5, LinkedIn and other famous social network platforms, Benevenuto Fabrício et al. ^[26] analyzed the behavioral law of 37,024 social network users, such as frequency of visiting social networks, activity types visited and the sequences of related activities. Golder Scott et al. ^[27] revealed the daily and weekly law that college students use social networks based on 362 million pieces of log information posted by 4.2 million users. With analysis of behavior patterns of 1.46 million users, Maia ^[28] pointed out that characteristics from social interactions are more effective for user clustering than that from individual user. Through analyzing the behavior of 80,000 users in Bebo, MySpace, Netlog, Tagged and other social networks, Gyarmati László et al. ^[29] discovered that the time that users spend on social network follows Weibull distribution and the duration of users' online sessions follows power-law distribution.

In this book, we describe the general law of the user behavior in online social networks from two aspects: user activities and their time pattern. From the aspect of user activities, we introduce the main activity categories in online social network activities and the transition law between those activity categories; from the aspect of time pattern, we introduce the law that users spend time on social network and the duration of users' online sessions.

1. Behavioral Law from the Aspect of User Activities

The activity is the basic unit of users' behavior in social networks, such as sending personal messages and browsing photos. An activity category contains several basic units. For example, photo category contains browsing photos, uploading photos, etc. In this section, based on the research of Benevenuto Fabrício et al. [26], we first introduce the method of data acquisition, and then the eight activity categories, and finally the transition law between activities.

1) Data acquisition

Benevenuto Fabrício et al. used Social Network Aggregator to collect users' clickstream. As an account management tool, Social Network Aggregator provides access to multiple social network accounts of the same user through common interface to realize the centralized management of social network accounts. Benevenuto Fabrício et al. collected clickstream data over a 12-day period from March 26 to April 6, 2009. The dataset contains all HTTP header information exchanged between users and the Social Network Aggregator, including time stamp, HTTP status, IP address of the user, etc. Data information after preprocessing is as shown in Table 5-15.

Table 5-15 Summary of the clickstream data

Online social network	Number of users	Number of sessions	number of requests
Orkut	36,309	57,927	787,276
Hi5	515	723	14,532
MySpace	115	119	542
LinkedIn	85	91	224
Total	37,024	58,860	802,574

Benevenuto Fabrício et al. obtained main activity categories in social network through analyzing clickstream log information, and then obtained the most common five activity categories in different social networks by analyzing the number of HTTP requests of each category of activity. They further analyzed the transition probability law from one activity category to another according to the sequence of user requests in clickstream.

2) Social activity category

As a user can conduct multiple kinds of activities in social network, Benevenuto Fabrício et al. enumerated all basic activity units of a user in social network, and then manually tagged each clickstream log entry with the appropriate activity category (e.g.,

friend invitation, browsing photos), and finally throw semantically similar activities into a category based on the webpage structure of online social network sites.

Benevenuto Fabrício et al. classified 41 identified activities into 8 categories as shown in Table 5-16. Global search enables users to capture other users' profiles, communities, and community topics in the whole Orkut sites. Scrapbook shows all text messages sent to a specific user. Scrapbook is different from personal messages or emails as it is public, i.e. all users having an Orkut account can read scraps of other users. Messages are a private way for communication and are available for each user. As a commentary function to scrapbook, testimonials show messages left to a given user by all his/her friends and are viewable to any user by default. Videos and photos include activities involving shared multimedia content. Profile & friends represents all activities related to profile management or browse of other users' profiles. Any user having an Orkut account can create communities, in which members can publish topics, inform others of major events, ask questions or play online games.

Table 5-16 Enumeration of all activities

Category	Description of activity	Number of users	Number of requests	Number of bytes (MB)
Search	1. Global search	2383	15409	287
Scrapbook	2.* Browse scraps	17753	147249	2740
	3. Write scraps	2307	7623	113
Messages	4.* Browse personal messages	931	3905	64
	5. Write personal messages	70	289	5
Testimonials	6.* Browse testimonials received from friends	1085	3402	57
	7. Write testimonials	911	4128	65
	8.* Browse testimonials written by oneself	540	1633	26
Videos	9.* Browse the list of favorite videos	494	2262	44
	10.* Browse a favorite video	390	862	13
Photos	11.* Browse the list of photo albums	8769	43743	871
	12.* Browse a photo album	8201	70329	2313
	13.* Browse photos	8176	122152	1147
	14.* Browse photos tagged by the user	1217	3004	47
	15.* Browse photo comments	355	842	16
	16.* Edit and organize photos	82	266	3
Profile & Friends	17.* Browse profiles	19984	149402	3534
	18.* Browse homepage	18868	92699	3866
	19.* Browse the list of friends	6364	50537	1032
	20. Manage friend invitations	1656	8517	144
	21.* Browse friend updates	1601	6644	200
	22.* Browse member communities	1455	6963	133
	23. Profile editing	1293	7054	369
	24.* Browse stars	361	1103	17
	25.* Browse user lists	126	626	9
	26. Manage user events	44	129	2

(To be continued)

Continued table

Category	Description of activity	Number of users	Number of requests	Number of bytes (MB)
Communities	27.* Browse a community	2109	8850	164
	28.* Browse a topic in a community	926	9454	143
	29. Join or leave communities	523	3043	43
	30.* Browse members in communities	415	3639	56
	31.* Browse the list of community topics	412	2066	38
	32. Participate in a community topic	227	1680	24
	33. Community management	105	682	12
	34. Post questionnaires in communities	99	360	6
	35.* Browse the list of communities	47	337	8
	36. Manage community invitations	20	63	1
	37. Community events	19	41	1
Others	38. Access applications	1092	4043	61
	39. User settings	403	2020	32
	40. Spam mail	48	150	2
	41. Account login and deletion	39	76	1
	Total	36309	787276	17.3GB

Note: Activities related to browsing behavior are marked with a (*) sign.

Table 5-16 shows users' interest in using various functions of Orkut. Browsing is the most common user behavior, which accounts for 92% of all user requests. For example, the number of users who ever browsed personal messages is 13 times larger than users who ever sent personal messages.

Benevenuto Fabricio obtained the most common five activity categories in the above four social network sites based on the number of HTTP requests as shown in Table 5-17. We can see from the table that Profile & Friends is the most popular activity across the above four social network sites. Users are likely to browse profiles of themselves or others, friend list, etc. in social network, which reflects the essence "sociality" of social network.

Table 5-17 Comparison of mainstream activity categories in different social network sites

	Orkut		MySpace		LinkedIn		Hi5	
Rank	Category	Share	Category	Share	Category	Share	Category	Share
1	Profile & Friends	41%	Profile & Friends	88%	Profile & Friends	51%	Profile & Friends	67%
2	Photos	31%	Messages	5%	Login	42%	Photos	18%
3	Scrapbook	20%	Photos	3%	Messages	4%	Comments	6%
4	Communities	4%	Login	3%	Search	2%	Login	4%
5	Search	2%	Communities	1%	Communities	<1%	Messages	3%

3) Transition law

Benevenuto Fabrício et al. established a first-order Markov chain of user activities

according to the sequence of user access activities. Figure 5-13 shows the transition probabilities among different activity categories. Nodes represent categories and directed edges represent the transition directions between two categories with weight indicating transition probabilities among different activities. Edges with probabilities smaller than 4% were removed. The sum of probabilities from one vertex to remaining vertexes is 1.

Benevenuto Fabrício found that most users began their sessions from the Profile & Friends, Scrapbook or Photos. This means that users will first conduct those activity categories upon their logging in social network sites. They also found that “self-loop” was very common. For instance, when a user participates in a community activity, the probability of still taking community in next activity is 0.82. Similarly, the probability of “self-loop” in photos activities reached 0.86. Except for “self-loop”, profile & friends was the most common next activity for current activities.

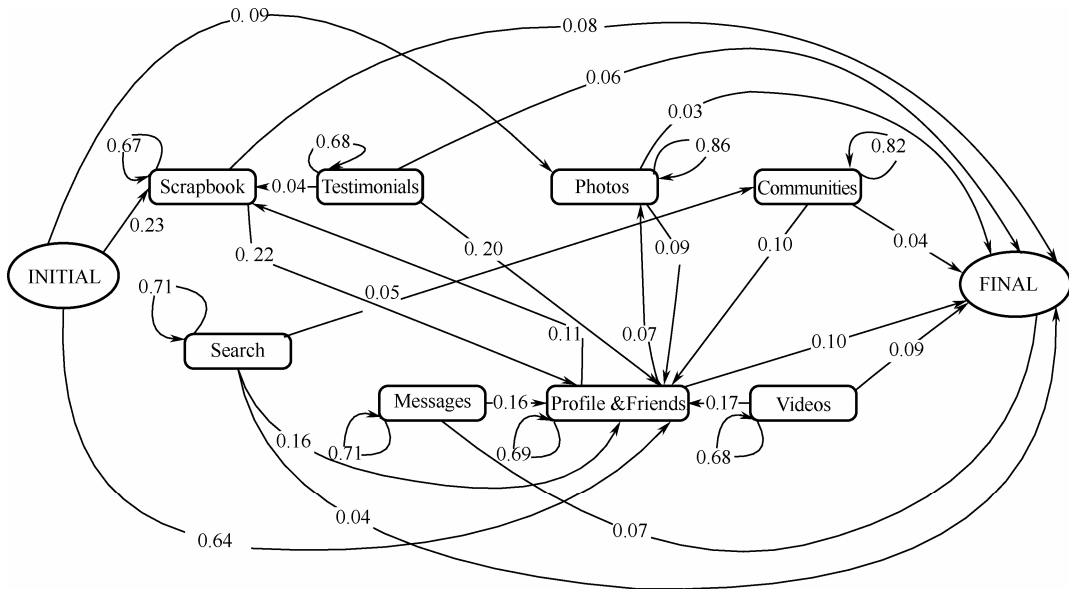


Figure 5-13 Transition law among different activity categories

2. From the Aspect of Time Pattern

The duration of online session refers to the time which users spend on browsing after logging in the sites. Session can be ended in two different ways: clicking the exit button or simply closing the Internet browser. In this section, based on the research of Gyarmati László et al. ^[29], we first introduce the collection method of behavior data from 80,000

users in the four social networks, and then present the law of time that the users have spent on social network, and finally introduce the duration law of users' online sessions in online social networks.

1) Data acquisition

As data cannot be directly obtained from the operators of the online social networks, Gyarmati László et al. captured the public part of users' profile pages in Bebo, MySpace, Netlog and Tagged. More than 500 PlanetLab nodes have been used to capture profile page information of more than 80,000 users from March 15 to May 2, 2009. Profile information refers to profile of user, status information posted, etc.

An appropriate monitoring time interval is very important as it is necessary to ceaselessly download the profile page of user during data acquisition. Under the condition of limited resources (only 500 nodes), short sampling time contributes to fine-grained dataset as more user profile pages will be downloaded; however, in this case the number of users monitored is limited. Monitoring more users needs larger time interval. According to practical experience, 1 minute was selected as monitoring interval. Gyarmati László et al. compared the changes between all downloaded profile pages of a given user to infer the duration of online state. For instance, if the current page content of a given user is different from that of 1 minute ago, it means that the user is still online. Finally, user's total online time in 6 weeks was obtained.

2) The law of time spent on social network

Figure 5-15(a) displays the cumulative distribution function of time that users have spent online during the period of monitoring. We note that the four social networks have similar characteristics (the figure of x axis is logarithmic), i.e. the majority of the users spent no more than a few thousand minutes online or less than an hour daily during the six-week period. Gyarmati László conducted fitting analysis and found that the measurement data and the curve of Weibull distribution overlap. This implies that the online time usage approximately follows Weibull distribution as shown in Figure 5-14(b).

3) The duration law of online sessions

Figure 5-15 displays the session duration distribution of user behavior in Tagged and MySpace (both x and y axes are logarithmic). We can observe that the session lengths follow power-law distributions. More specifically, the figure shows the sum of two power-law distributions and the jump is their boundary point. The change between the distributions is mainly due to the session timeout settings of the servers (both Tagged and MySpace session have a 20-minute timeout period). Timeout setting means that the

session remains active for some time after users end the session by simply closing the browser. Therefore, the session ended by clicking exit button can be shorter than that by closing the browser. Ending session by exiting from social network sites constitutes the first power-law distribution, while the other distribution is the sum of the two ways of exiting.

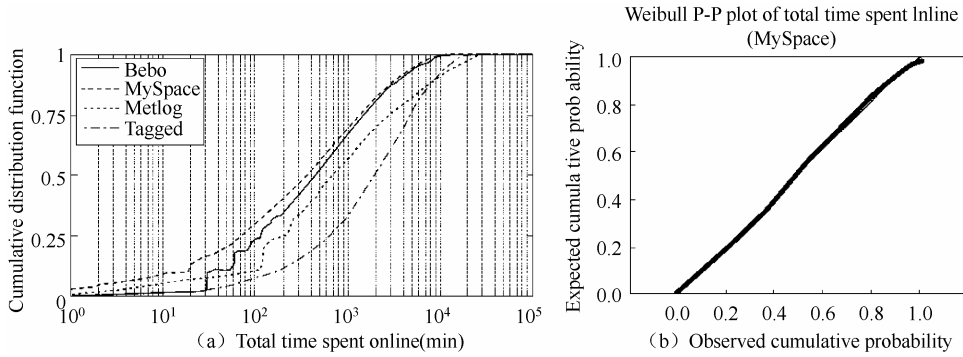


Figure 5-14 Distribution function of usage time

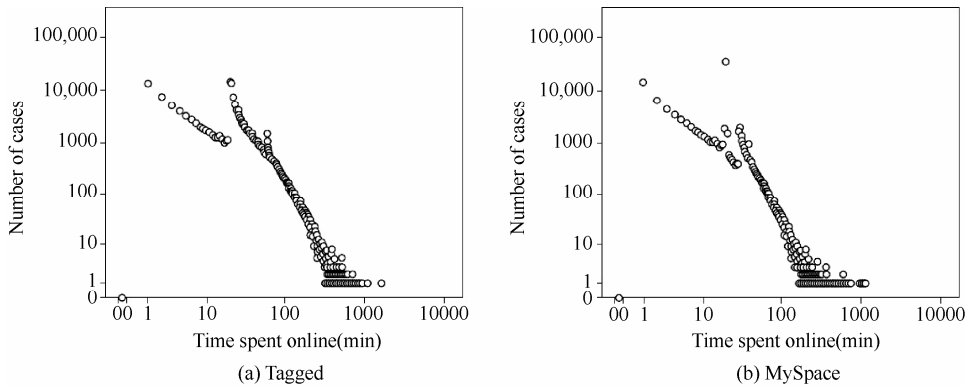


Figure 5-15 Session distribution

5.3.2 Behavior of Content Generation

With the rapid development of Web 2.0 and the rapid growth of social network applications, the interactivity and We-Media property of the Internet become increasingly prominent. Compared with Web 1.0, the users in the age of Web 2.0 become more autonomous and interactive. The online users are no longer media audiences passively accepting information, but actively participate in the production, release, diffusion and

sharing of the information, which produces a large amount of User-Generated Content (UGC). For example, the cumulative number of Tweets generated per day is 250 million in October 2011. Only a few months later in March 2012, the number surprisingly reaches 310 million. In the face of such a huge amount of User-Generated Content, research on the user generation behavior is of prime importance as companies cannot make good use of social media and develop effective social media marketing strategies without full understanding of the law of UGC.

Shriver Scott et al. ^[32] found a significant causal relationship between UGC and user network structure, i.e. UGC is affected by user network structure. Furthermore, Toubia Olivier et al. ^[33] pointed out that the intrinsic and image-related utility influenced UGC. Depending on the creation motivation, the increase of followers can stimulate the content generation, and also reduce the users' enthusiasm for generating new content. Previous research has paid attention to the disclosure of users' personal information, including literary quality, entertainment interest, political tendency, etc. Although big risk lies in disclosing personal information, research has shown that users are still enthusiastic about sharing their personal information. Furthermore, the disclosure preference of users (topic selection) is also widely noted. In-depth research on topic selection behavior in UGC is of great significance for the personalized recommendation system. For example, Wang Yi-Chia et al. ^[34] pointed out that women tended to generate more personal topics (e.g., family matters), while men discussed more public topics (e.g., politics and sports). Many scholars, especially psychology scholars, have probed into the expression behavior of UGC. Yarkoni Tal ^[35] analyzed the relationships between users' personality traits and 100 thousand words posted by blog users, and obtained many meaningful conclusions. For example, users with high sense of responsibility rarely used negative words. Qiu Lin et al. ^[36] analyzed all Tweets published by 142 users, and found that their expression behaviors follow some regulations. For instance, extravert users tend to use positive sentiment-specific words.

Based on the above analysis, we first expound the motivation for content generation, and then introduce the self-presentation behavior in content generation. Furthermore, we introduce the topic preference during users' self-presentation, and finally introduce an important form of content generation: language expression behavior.

1. Motivation for Content Generation

In this subsection, based on the research of Toubia Olivier et al. ^[33], we introduce the impact of the increase of followers on content generation behavior under different creation motivation from two aspects of intrinsic and image-related utility. Specifically, we first

expound the theoretical basis, and then introduce the experiment data and method, and finally summarize the results.

1) Theoretical basis

Utility refers to a measure of consumer satisfaction, which is obtained by consumption or enjoyment. In this book, intrinsic utility indicates that online users, for its intrinsic satisfactions rather than for other reasons, get direct utility from publishing content, i.e. users can get psychological comfort when they post content to their followers. Image-related utility assumes content generation of users is motivated by the perceptions of others, such as pursuing prestige. Image-related utility is not limited to users' appearance. Instead, image-related utility contains the sense of self-worth and social acceptance. Image-related utility is directly derived from the number of followers, and posting more content may attract more followers.

If users contribute content to Twitter for intrinsic utility, the utility function of user post behavior is concave not decreasing along with the increase of the number of followers, i.e. users are likely to increase their posting activities as they have additional followers. The image-related utility is derived from the number of followers, i.e., the more followers, the higher utility. Posting content can increase the number of followers in the future, and then influence the image-related utility. In contrast to intrinsic utility, the incremental image-related utility generated by posting content on a given day will be obtained in the future, and depends on the number of additional followers in the future. If incremental image-related utility generated by increase of followers decreases, users will be negative in posting new content.

In summary, we can give opposite predictions on intrinsic utility and image-related utility according to users' response to an increase in the number of followers. If users are motivated by the intrinsic utility, more followers will contribute to an increase in posting activities. On the other hand, if users' utility result from attracting more followers and they post content for gaining additional followers, the motivation to post content will decrease as the increase in the number of followers due to utility diminishing law.

2) Data and experiment method

Toubia Olivier et al. randomly selected a collection of 2,493 non-commercial Twitter users from the initial database with nearly 3 million users. They collected three variables daily for each user: the number of followers, the number of users followed and the total number of Tweets since account creation; 1,335 users were classified as active users based on certain criteria. The experimental period was 160 days, and 100 users were selected

from active users as experimental group. In order to research the influence of the number of followers, Toubia et al. gradually added 100 public followers to the users in the experimental group over a 50-day period. For example, they started by adding one follower per day to each user for four days, then added two followers per day for another four days and so on until the 100 followers were added for each user. Concrete operation process is as shown in Figure 5-16.

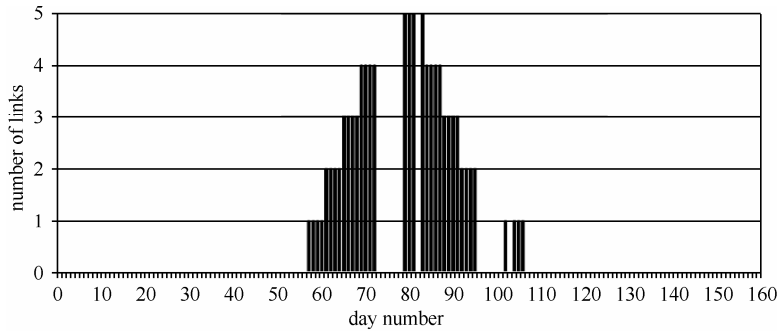


Figure 5-16 The process of adding followers

3) Research conclusions

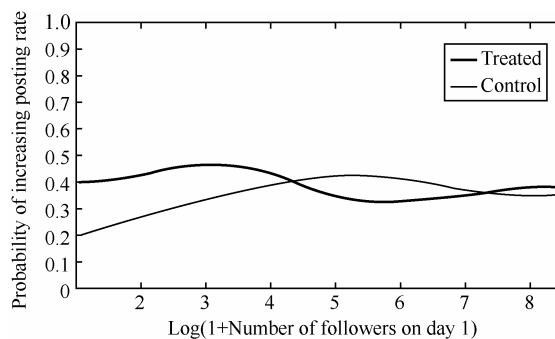
Based on the above analysis, when the number of followers increases, an increase (decrease) in posting content conforms to intrinsic utility (image-related utility). After the intervention (adding followers to selected users), users in experimental group are more enthusiastic in posting content than that of controlled users. Specifically, 40.82% of users in experimental group became more active compared with 34.19% of users in controlled group. However, the difference is not statistically significant. Therefore, the intervention did not have a significant influence on posting activities.

Furthermore, Toubia Olivier et al. researched the influence of intervention on users' posting activities when initial number of followers is different. There are two considerations. First, they consider intrinsic and image-related utility as the function of a user's number of followers. Therefore, the behavior of a user with few followers may be more consistent with one utility, while it may become more consistent with the other utility as the number of followers increases. Second, the relative importance between image-related and intrinsic utility may be different for different users, which can be reflected in the number of followers. For example, users having more followers may be more concerned about intrinsic utility.

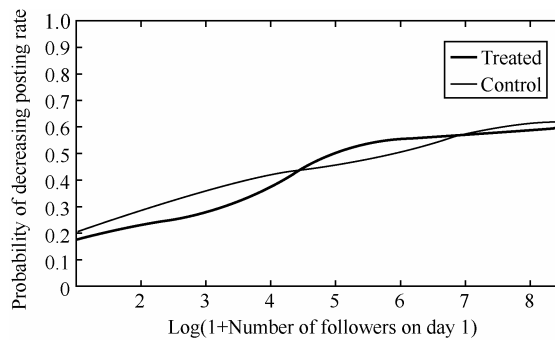
Figures 5-17 and 5-18 respectively plot the probabilities that users increase and decrease their posting activities before and after the intervention. The x axis refers to the log of users' initial number of followers. Toubia Olivier et al. found that selected users with few initial followers tended to post more content after intervention. However, treated users with many initial followers tended to decrease their posting frequency.

In order to statistically compare the influence of intervention on posting behavior of users with different initial number of followers, Toubia Olivier et al. partitioned the selected users into five subgroups as shown in Table 5-18. Compared with controlled group, users in the second experimental subgroup were more likely to increase their posting frequency. However, the users in the fourth experimental subgroup acted in an opposite manner. The differences in other subgroups had no statistical significance.

In fact, it is not surprising to see the consistent performance of the first and fifth subgroups. Users in the first subgroup have few followers and seldom visit the Twitter platform; users in the fifth subgroup have many followers and were not sensitive to the addition of 100 followers over the 50-day period. Therefore, increase of followers has no significant influence on the first and fifth subgroups. Behavior of users in the second subgroup conforms to the intrinsic utility, and, as discussed above, intrinsic utility derived from having more followers prompt users to post more content. The performance of users in the fourth subgroup conforms image-related utility, and, as discussed above, the increase of follower decreases the posting frequency as the image-related utility decreases with the increase of followers (followers attracted by posting content). Users in the third subgroup are influenced by both intrinsic and image-related utility, which counteract each other.



Figures 5-17 Probability of increasing posting activities before and after the intervention



Figures 5-18 Probability of decreasing posting activities before and after the intervention

Table 5-18 The partition of users

Subgroup	Range of number of followers	Median number of followers	Average number of followers
1	0 ~ 12	7	6.499
2	13 ~ 26	19	18.941
3	27 ~ 61	39.5	40.988
4	62 ~ 245	109	125.550
5	246 ~ 18,940	704	1,378.949

2. Self-presentation Behavior

In this subsection, we first introduce the general rule of self-presentation behavior, and then the effect of gender on self-presentation by photos.

1) General rule of self-presentation behavior

Boyle Kris et al. ^[37] found that, compared with directly listing their names in the page (59.7%), users tended to present themselves by uploading photo (75.7%), creating a motto (74.7%), and listing personal interests (79.9%). Surveys showed that users were reluctant to create blogs (33.3%), post videos (23.9%), and upload slideshows (23.9%). Only 5% users posted videos of themselves or their friends and family. Comparatively speaking, MySpace users preferred to post slideshows to record their friends and family (11.6%) or their things with friends (4.8%).

Users' disclosure levels of individual information differ greatly. Almost all respondents listed their relationship status (99.2%), hometown (97%) and postcode (99.2%). Most users disclosed their race (83.3%), sexual orientation (82.5%), whether they plan to have a baby (79.9%), education (79.9%). Although 79.9% users listed personal interests, relatively few users were willing to disclose their favorite music (69.1%), movies (60.2%), television shows (59.4%), books (53.0%), or idols (57.8%).

Few users were willing to reveal their income (19.1%) or groups they belong to (19%).

2) Self-presentation by photos

Based on the research of Tifferet Sigal et al. ^[46], we researched the influence of gender on the selection of profile and cover photos. Profile photo refers to head portrait in the homepage. Males' photos highlighted status (e.g., dress formally) and adventure (e.g., outdoor activities), while females' photos highlighted family relations (e.g., family photos) and emotional expressions (e.g., eye contact, smile intensity). Cover photo refers to background picture at the top of the homepage. Gender had no significant influence on the selection of cover photo. The only difference is that females were more likely to show family photos on the cover.

Tifferet Sigal et al. pointed out that Facebook users prefer to present themselves with their profile photos, and nearly 60% users use their own photos as the profile photos. Moreover, users used their cover photos to extend their presentation, e.g., introducing additional aspects of the self.

3. Topic Preferences

In this subsection, based on the research of Wang (2013), we first introduce methods for data acquisition and preprocessing, and then construct topic model by Latent Dirichlet Allocation (LDA), and finally discuss the relation between gender and topic selection behavior of users.

1) Data acquisition and preprocessing

Wang randomly selected one million status updates posted by American Facebook users in June 2012. For each status update, Wang analyzed its metadata including posting time, number of viewers, number of comments and likes within three days after posting. Wang also collected demographic information of each user, including gender, age, number of friends and number of days after registration. The preprocessing of collected information is divided into three steps. First, tag all status updates with OpenNLP ^[38]. Second, process stem with Porter stemmer to remove all punctuations, URLs, email addresses and tags. Third, represent all updates as the combination of unigrams and bigrams. For example, the sentence "I like this photo" was represented as <I, like, this, photo, I like, like this, this photo>. We termed the first four as unigrams (single word) and the last three as bigrams (word pair).

Of all terms in the one million status updates, 71% of unigrams appeared only once, and the top 500 most frequent unigrams accounted for 55% of the whole corpus. For instance, 10% of status updates contained “love” which was the most frequent term. High frequent terms always co-occurred with different terms, and thus were not useful in topic modeling. Similarly, the low frequent terms were also not useful because they almost did not co-occur with other terms. Therefore, Wang removed high (occurred in more than 0.5% of the corpus) and low (occurred in less than 0.01% of the corpus) frequent terms to reduce noise and vocabulary size. Furthermore, they excluded 500 unigrams based on a stoplist and bigrams were also removed if both two words were stopwords. Nearly 50% of the updates had unigrams or bigrams less than 8 after pruning those which were too short for successful topic modeling.

2) Topic modeling

Wang et al. applied Latent Dirichlet Allocation (LDA) model to identify topics from the remaining 521,636 status updates. LDA is a statistical generative method, which is often used to discover hidden topics and the words associated with each topic. It can analyze a large number of unlabeled documents by clustering words that frequently co-occur. Wang et al. tried to run LDA models with 10, 30, 50, 60 and 100 topics. The comparison showed that the results were the most interpretable when topic number was set to 50. Wang et al. generated a dictionary for each topic based on the 500 terms that were most associated with that topic. Two experts familiar with social network content manually examined each dictionary. Through the process of removing and merging, they finally got 25 standard topics as shown in table 5-19. We consider a status update belongs to a topic if it has at least 3 words thereof.

Table 5-19 Overview of topic categories

Topic	Sample term
Sleep	last night, this morning, wake up, sleep, bed
Food	lunch, cook, coffee, beer, chicken, cake, ice cream
Clothing	supermarket, line, wear, store, cloth, dress, bag
House	door, my house, cat, street, box, window, floor
Weather/Travel	road, weather, cold, city, air, town, fly, storm
Family fun	great day, time, kid, swim, cousin, evening party
Girlfriend/Boyfriend	best friend, girl, red, boyfriend, my favorite
Birthday	I love, love you, my baby, happy birthday, birthday
Father's Day	happy father, father's day, children, my dad

(To be continued)

Continued table

Topic	Sample term
Sports	beat, fan, ball, ring, Miami
Politics	national, country, the U.S., president, vote, law
Love	my heart, give, strong, love me, happy
Thankfulness	thank you, visit, my family, thankfulness, Thank God
Anticipation	wait for, celebrate, can't wait, wow, until
Asking for support/pray	My friend, worry about, help me, right now, pray, support
Medical	drop, doctor, hospital, test, blood, stress
Memorial	miss, memory, everyday, peace, grandma
Negativity about people	say, judge, waste, some people, rubbish
Complaining	hate, guess, tired, talk
Deep thoughts	idea, success, human, create, symbol, goal, universe
Christianity	God, faith, church, Christianity, spirit
Religious imagery	death, man, star, birth, earth, sun
Slang	luv, knw
Work	to work, working, boss, colleague

Note: Each above expression contains two or less words.

3) Conclusions

To research the influence of gender and age on topics selected by users in social network, Wang et al. recalculated the popularity of each topic for men, women, boys and girls. The popularity of a given topic refers to the percentage of updates that contains that topic. Figure 5-19 plots the differences in topic preference for women and men at the age of 25 and above. Women prefer to discuss topics related to personal details (e.g., Father's Day, family fan and birthday), while men prefer to discuss sports and some abstract concepts (e.g., politics and deep thoughts). Despite the differences in language format and audience in social networks, this finding conforms to previous research results on the influence of gender on face-to-face communication and blogs.

Gender has no significant influence on teenagers' selection of topics. Figure 5-20 shows the topic selection preference^[39] of girls and boys at the age of 13 ~ 16. Compared with adults, teenagers have no significant gender difference in selection of topics. For example, complaining, girlfriend/boyfriend, and slang were the most popular topics for both teen boys and girls.

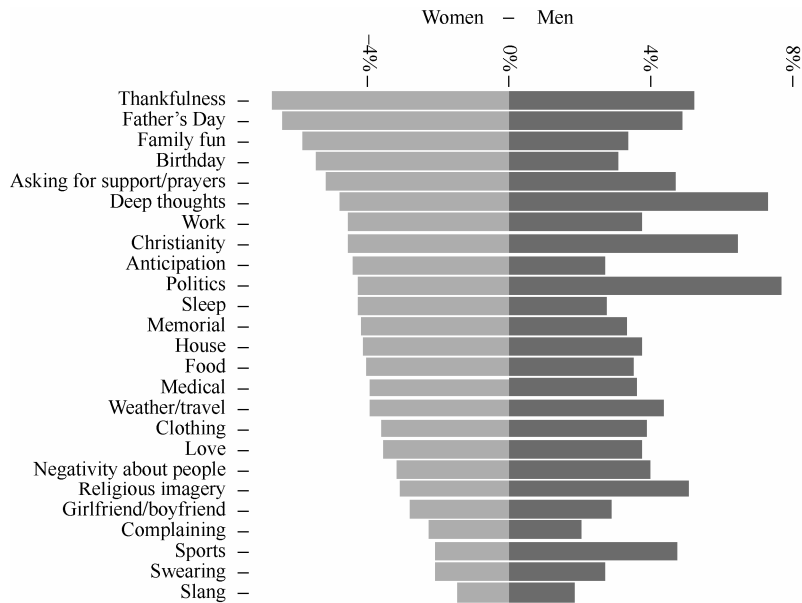


Figure 5-19 The influence of gender on topic preferences (at the age of 25 and above)

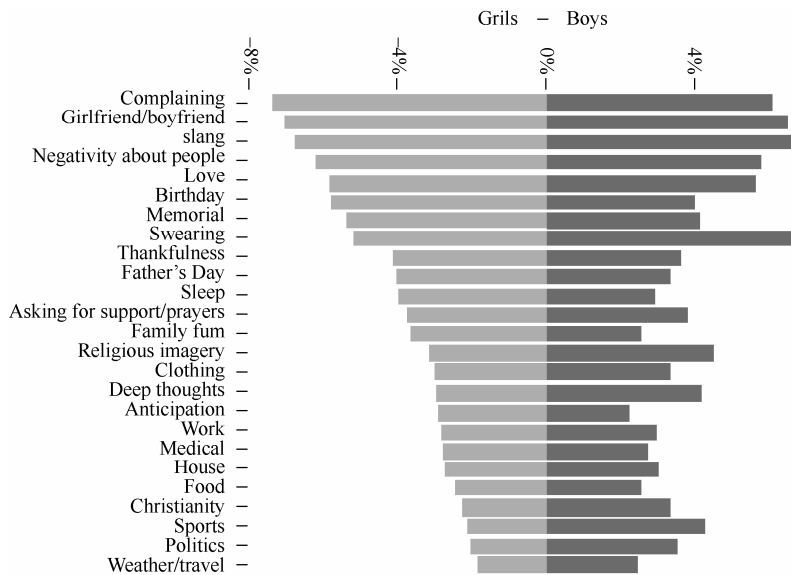


Figure 5-20 The influence of gender on topic preferences (teenagers)

4. Language Expression Behavior

Previous research on expression behavior was mainly based on the LIWC^[40], which can extract 81 characteristics from a passage and compute the proportion of each

characteristic. Many research achievements about the relationships between personality traits and language expression behavior in social networks have been obtained by a large number of scholars. Social network platforms involve blog, Facebook, Twitter, etc. Main results were shown in Table 5-20.

Table 5-20 The relationships between personality traits and language expression behavior

Author(s)	Platform	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Golbeck ^[41]	Facebook	Perceptual process (-), work	Emotion, positive emotion, physiological process	Bad language (-), social process, human, perceptual process (-), visual (-)	Anxiety, ingestion	Money (-)
Nguyen ^[42]	Blog	Word consisting of 6 letters, leisure, number, money, perceptual process, bad language (-), nonfluency (-), health (-), negation (-)				
Nowson ^[43]	Blog	Occupation (-), achievement (-), discrepancy (-), school (-), human, social process	Discrepancy (-), word consisting of 6 letters, article, negation (-)	Death (-)	Discrepancy, work, anxiety, future tense verb, eating, human (-)	Word consisting of 6 letters, positive emotion, school (-), occupation (-), modifier, inclusive, preposition
Yarkoni ^[35]	Blog	1st person plural pronoun, 2nd person pronoun, number(-), positive emotion, causation (-), inhibition (-), tentative (-), certainty, hearing, social process, friend, family, human, inclusive, occupation (-), work (-), achievement (-), music, religion, sexuality	Pronoun, 1st person plural pronoun, 1st person pronoun, numbers, positive emotion, negative emotion (-), anger (-), causation (-), seeing, social process, friend, family, time, past tense verb, space, inclusive, motion, leisure, home, music, money (-), death (-), sexuality, sleep, bad language (-)	Negation (-), negative emotion (-), anger (-), sadness (-), cognitive process (-), causation(-), discrepancy (-), tentative (-), certainty (-), hearing (-), human (-), time, exclusive (-), achievement, music (-), death (-), bad language (-)	1st person singular pronoun, 2nd person pronoun (-), negation, article (-), negative emotion, anxiety, anger, cognitive process, causation, discrepancy, tentative, certainty, friend (-), space (-), exclusive, sleep, bad language	Pronoun (-), 1st person singular pronoun (-), 1st person plural pronoun (-), 1st person pronoun (-), 2nd person pronoun (-), negation (-), assent (-), article, preposition, number (-), affect (-), positive emotion (-), cognitive process (-), discrepancy (-), social process (-), family (-), human (-), time (-), past tense verb (-), present tense verb (-), future tense verb

Some of the conclusions were consistent with each other and could be mutually confirmed, but more showed the differences and contradictions. Therefore, the relationships between personality traits and language expression behavior need further verification. For example, Golbeck Jennifer et al. ^[41] found extravert users were more likely to use words about work, while Yarkoni et al. ^[35] obtained opposite results.

5.3.3 Behavior of Content Consumption

Tens of thousands of users visit social network (e.g., Flickr, Facebook) every day to share their photos, videos, mood, etc., while others satisfy their requirements by searching, viewing or commenting such information. Benevenuto Fabrício et al. ^[26] found that browsing behavior accounted for 92% of all user activities by analyzing users' clickstream, such as browsing others' profile page, watching the videos shared by others, or browsing others' photo albums. By analyzing such "silent" user behavior, we can get a more accurate and comprehensive view of the online social network workload. Most user behavior is passive, i.e. consuming the content created by others. A large number of scholars focused on this topic from front and side aspects, and have obtained many meaningful results. For example, we can know, from the side aspect, the users' consumption preferences by analyzing the question type asked in social networks.

Based on the access logs in Flickr, a site based on photo sharing, Van Zwol et al. ^[45] analyzed the users' browsing behavior from temporal, social and spatial dimensions and answered the time, reason and location of users' browse behavior. They found that users were able to browse new photos within hours after being uploaded, and most browse behaviors occur within the first two days. With the continuous development of social networks and the explosion growth of information therein, social networks become a more and more important information repository. For example, Facebook launched a social graph search tool called Graph Search in July 2013, which focuses on helping users search content by relational network compared with the traditional network research featuring key words and phrases. Since then, users used social networks to obtain information in addition to entertainment. There are two typical ways to obtain information. One is the passive consumption behavior, and the other is active consumption behavior, which refers to the phenomenon that people ask

questions by social network status messages instead of traditional search engine (e.g., Google). A large number of scholars begin to pay close attention to this area. By analyzing 624 users in social network, Morris Meredith Ringel et al. ^[44] found that more than half of users had asked questions through social networks, and there were significant differences in question type and the way of asking.

Based on the above discussion, we intend to introduce the behavior of content consumption in social network from two aspects: passive and active consumption. We introduce the browsing behavior law from the passive aspect and information acquisition behavior law from the active aspect.

1. Passive Consumption—Browsing Behavior

Based on the research of Van Zwol Roelof et al. ^[45], we first introduce the data acquisition and processing, and then research the browsing behavior from temporal and spatial dimensions wherein temporal dimension focuses on tracking the change of browsing behavior over time, and spatial dimension aims to investigate the geographic distribution of browsing behavior.

1) Data acquisition and processing

Van Zwol Roelof et al. tracked the number of views of 1.83 million photos for the duration of 60 days. The observation window was a 50-day period. To be exact, if a photo was uploaded on day 3, then the authors took the number of views for that photo into account up to day 53. Van Zwol Roelof et al. only considered those page views that were explicitly oriented on viewing a single photo. By using Yahoo!'s IP address to geographic location conversion service, the authors could inference the origination of the photo view requests. Figure 5-21 shows the log-log distribution of photo views. The *X* axis refers to the 1.83 million photos, which were ranked by the number of views. The *Y* axis represents the total number of views per photo. This distribution could be fitted accurately by power-law distribution. The distribution of the photo views had high slope, i.e. a small fraction of photos received most views.

Based on the number of views, Van Zwol Roelof et al. classified all the photos into ten groups with equal number of photos and computed the total views for each group. As shown in Table 5-21, the first group (0 ~ 10%) contained the photos that were most frequently browsed, while the groups for 50% and above contained the photos that were only viewed once.

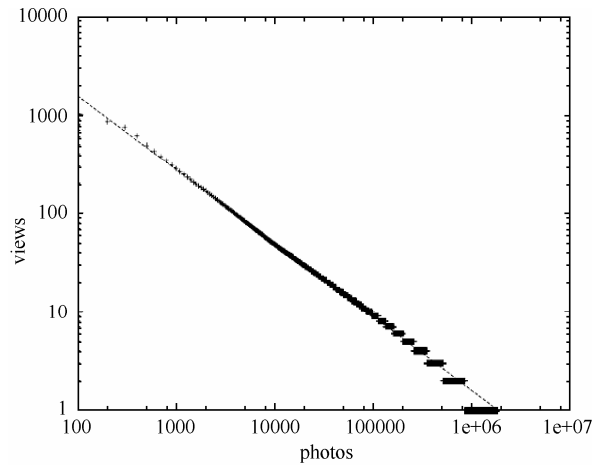


Figure 5-21 The distribution of photo views

Table 5-21 Photo classification

Group	Number of views	Group	Number of views
0% ~ 10%	3,802,875	50% ~ 60%	182,856
10% ~ 20%	812,131	60% ~ 70%	182,857
20% ~ 30%	515,532	70% ~ 80%	182,856
30% ~ 40%	365,712	80% ~ 90%	182,857
40% ~ 50%	312,270	90% ~ 100%	182,857

2) Temporal dimension

Figure 5-22 shows the total number of photo views after photos being discovered. The number of views for the first group after 3 hours approximated that for the second group after 50 days. Moreover, popular photos had 45% of their total views within the first 48 hours. More detailed information about Figure 5-22 could be found in Table 5-22, which provided the mean and standard deviation of the total views over time. The standard deviation of the first group grew significantly over time, while the second group showed little change.

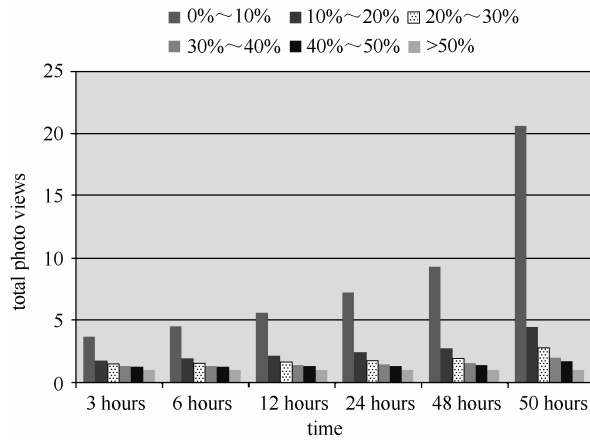


Figure 5-22 Change law of photo views

Table 5-22 Change law of photo views

	3 hours		6 hours		12 hours	
Group	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
0%~10%	3.63	8.25	4.44	12.51	5.55	18.66
10%~20%	1.77	0.97	1.92	1.05	2.11	1.12
20%~30%	1.47	0.67	1.54	0.7	1.62	0.73
30%~40%	1.3	0.46	1.33	0.47	1.36	0.48
40%~50%	1.25	0.43	1.27	0.44	1.3	0.46
>50%	1	0	1	0	1	0
	24 hours		48 hours		50 hours	
Group	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
0%~10%	7.24	26.5	9.28	37.6	20.6	87.7
10%~20%	2.43	1.22	2.75	1.28	4.4	0.7
20%~30%	1.77	0.77	1.92	0.79	2.8	0.4
30%~40%	1.44	0.5	1.52	0.5	2	0
40%~50%	1.35	0.48	1.39	0.49	1.7	0.45
>50%	1	0	1	0	1	0

3) Spatial dimension

Table 5-23 showed the top 5 most frequently viewed groups and we only took the 10% of photos for each group. The second column of the table referred to the average geographic diffusion of the photo views in a group. The decreasing mean indicated that the geographic diffusion generally focused on a specific geographic location when the popularity of a photo decreased. The third column indicated the standard deviation of the

geographic diffusion with the smallest standard deviation for the first group and relatively larger values for the other groups. One explanation for this browsing behavior was that, in the groups >20%, the location where users browse photos were near to the uploading location, i.e. users preferred to browse the photos near their locations. The fourth column represented mean views of each photo in each group.

Table 5-23 User's interest related to space

Group	Mean	Standard deviation	location
0%~10%	28.61	10.44	740.2
10%~20%	18.84	19.29	26.4
20%~30%	13.23	19.83	5.28
30%~40%	9.54	19.57	2
40%~50%	9.03	15.99	1.69

2. Active consumption—information acquisition behavior

For active user information acquisition behavior, we first compare the differences in information acquisition between search engine and online social network, and then introduce the consumption preference when users acquire information in social networks.

1) Comparison of information acquisition behavior between search engine and social network

We compared the differences between social and non-social search based on the research of Morris Meredith Ringel et al. ^[47]. Social search meant that users obtained information by asking questions through status updates of social networks (e.g., Facebook), and then waited for their friends to provide answers. Non-social search meant to obtain answers through traditional search engine.

Morris Meredith Ringel et al. compared the difference in solving complex information acquisition task between social and non-social search through a lab study. 12 participants (4 females), who were all Microsoft employees and aged between 23 and 42 years old (mean age = 31.9), took part in this experiment. 5 participants called themselves as expert searchers, and the remaining average. All participants were asked to have at least 50 friends on Facebook to ensure that their social network was large enough to provide answers to their questions. Network size ranged from 50 to 743 (mean size = 260.3).

All the questions (e.g., shopping and travel) were prepared by participants before experiment. When the experiment began, the participants posted a question through status update in social network, and then independently search answers by search engine.

Participants could determine to end the search by themselves when they felt satisfied with what they had found, and each query, URL and corresponding timestamp are saved by a client plug-in. When participants finished the search task by search engine, Morris Meredith Ringel et al. asked them to check their social network and capture a screenshot of all responses and timestamps to their questions uploaded. Three days later, the participants further checked their social network and captured a screenshot of the new responses.

As shown in Table 5-24, Morris Meredith Ringel et al. compared the social and non-social search by non-parametric Wilcoxon tests. Participants spent an average of 30.3 minutes on the Web to search answers. On average, they submitted 6.5 queries and browsed 35.4 pages from 12.3 distinct, non-search sites. Through social search, they received an average of 1.4 responses while five participants did not get any responses. In the following three days, participants received an average of 4.1 responses. The number of responses received by each user ranged from 0 to 20 with two participants did not get any responses. Of the ten participants, the minimum time for receiving the first response was 5 minutes. Time to the first response was negatively associated with the number of friends.

After comparing answers from social search and non-social search, 11 participants (91.7%) were more satisfied with the answers from search engine and its performance seemed more to better meet users' expectations. An important reason that participants preferred search engine was that they could get answers more quickly. Nevertheless, the benefits of social search were apparent. 8 participants (66.7%) had asked questions through social network before and they thought that their friends may provide more customized answers since they knew additional background information (e.g., personality, preferences) of the participants.

Table 5-24 Overview of social search results

Task	Network size	Initial responses	Total responses	Time to the first response (minute)	Search time (minute)
I want buy a new phone... Any suggestions?	466	3	20	15	38
Any tips for tiling a kitchen backsplash?	231	3	7	8	29
Does anyone know how to stop an in-car nav system from constantly rebooting?	275	2	2	19	46
Does anyone know how to train for half marathon?					
Does anyone know any good vegetarian recipes?	50	0	0	N/A	21
So...after getting the PMP, what else can I do to keep up their development?	401	1	10	36	36

(To be continued)

Continued table

Task	Network size	Initial responses	Total responses	Time to the first response (minute)	Search time (minute)
Should I buy iPod as a gift?	104	1	3	7	32
I want to move away from LiveSpace for storing and sharing photos... Any recommendations?	206	0	5	184	12
Can I flee from Seattle winter by taking a trip to New Zealand?	240	0	5	77	31
Any recommendations on restaurants and activities in Cancun	143	2	2	5	49
What are the must see attractions in Disneyland?	743	5	10	8	22
Any recommendations on a good terminal or high end TV?	169	0	0	N/A	34

2) Consumption preference

When acquiring information, social search had other obvious superiority despite its slower search speed than that of search engine, i.e. social network can provide customized answers based on users' preference. Therefore, it is of important value to study social search. Based on the research of Morris Meredith Ringel et al. ^[44], this subsection introduces the consumption preferences of content in social network by survey method.

624 participants (1/4 female) were involved in the surveys. Since social networks were heavily used by college students, it is ensured that the majority of participants (72.2%) were full-time employees and the remaining were summer interns. This partition made the samples more representative. To eliminate the intervention of age, it is ensured that 28% of the participants were aged 18 ~ 24, 40.1% aged 26 ~ 35, 25.5% aged 36 ~ 45, and only 6.1% aged 46 and above. Besides basic information, all the participants were asked to report a series of questions related to question asking and answering behaviors. For example, whether participants have ever used social networks to ask questions. If they had done so, they were further asked about a set of questions about the frequency of such behavior, the type and the topic of the questions.

Morris conducted non-parametric tests to detect the significance of differences in the distribution of question topic and type over gender and age. Morris first explored the question type. Question type referred to the nature of the question. For example, whether the question belonged to recommendation or invitation. Table 5-25 listed the main categories and popularity of the different question types. We can see that the most prevalent question types were for recommendation and opinion. Both the two types asked a user's

friends to provide subjective information. An opinion question usually asked for a rating of a specific item, while a recommendation question was an open-ended request for suggestions. For example, a user may hope that their friends can suggest a cheap and fine mobile phone.

Table 5-25 Question types and the corresponding popularity

Question Type	Percent	Example
Recommendation	29%	Building a new playlist...any ideas for good running songs?
Opinion	22%	I am wondering if I should buy the ice cream maker?
Factual knowledge	17%	Anyone know a way to put Excel charts into LaTeX?
Rhetorical	14%	Is there anything in life you're afraid you won't achieve?
Invitation	9%	Who wants to go to Navya Lounge this evening?
Favor	4%	Does anyone need a babysitter tonight?
Social connection	3%	I am hiring in my team. Do you know anyone who would be interested?
Offer	1%	Could any of my friends use boys size 4 jeans?

Second, in addition to question type, Morris also introduced the popular question topics. Topic (e.g., technology, music) referred to the subject matter of the question. Table 5-26 showed the categories and popularity of question topics, and illustrated each topic using an example. The most popular question topic was technology, including computer hardware, software, programming, mobile phones, etc. Entertainment questions were also popular, which included movies, television, arts, music, etc.

Morris explored how the participants' demographics and social network use correlated to the types and topics of questions. The details were shown in Table 5-27. There was no significant gender differences in the types of questions asked. However, gender had a larger impact on question topic. More specifically, men asked a higher proportion of technology questions, while women preferred to ask family-related questions. Age was correlated with the type of questions asked. Compared with older people, younger participants preferred to ask invitation questions. In contrast, older participants preferred to be recommend by others. Age had no significant impact on the topic of questions asked. Furthermore, Morris found that the participants were more likely to ask technology questions on Twitter, while they were more likely to ask questions about entertainment and home & family on Facebook. Participants, who infrequently updated their status, were more likely to ask questions associated with rare events or special occurrences, such as travel and health.

Table 5-26 Question topics and the corresponding popularity

Question Topic	Percent	Example
Technology	29%	Anyone know whether WoW works on Windows 7?
Entertainment	17%	Was seeing Up in the theater worth the money?
Home & Family	12%	So what's the going rate for the tooth fairy?
Professional	11%	Which university is better for Masters? Cornell or Georgia Tech?
Places	8%	Planning a trip to Whistler in the offseason. Recommendation on sites to see?
Restaurant	6%	Hanging in Ballard tonight. Dinner recs?
Current events	5%	What is your opinion on the recent proposition that was passed in California?
Shopping	5%	What's a good Mother's Day gift?
Ethics & Philosophy	2%	What would you do if you had a week to live?

Table 5-27 The impact of participants' demographics and social network use on the type and topic of questions

		Gender		Age				Network size		Frequency of Use	
		Male	Female	18~25	26~35	36~45	46~55	Facebook	Twitter	Infrequent	Frequent
Total		157	77	51	93	71	15	126	49	205	29
Question Type	Opinion	23.6%	22.1%	21.6%	18.3%	28.2%	33.3%	18.3%	30.6%	20.7%	23.4%
	Recommendation	31.2%	29.9%	13.7%	35.5%	38.0%	26.7%	31.0%	28.6%	24.1%	31.7%
	Factual Knowledge	15.3%	13.0%	9.8%	22.6%	9.9%	6.7%	11.1%	26.5%	13.8%	14.7%
	Rhetorical	8.9%	16.9%	19.6%	6.5%	11.3%	20.0%	14.3%	6.1%	13.8%	11.2%
	Invitation	10.8%	10.4%	23.5%	10.8%	2.8%	0.0%	15.1%	4.1%	24.1%	8.8%
Question Topic	Technology	35.0%	22.1%	33.3%	26.9%	32.4%	40.0%	15.9%	61.2%	24.1%	31.7%
	Entertainment	17.8%	19.5%	19.6%	24.7%	11.3%	6.7%	24.6%	6.1%	24.1%	17.6%
	Home & Family	8.3%	19.5%	7.8%	14.0%	14.1%	6.7%	19.0%	0.0%	13.8%	11.7%
	Professional	10.8%	9.1%	7.8%	9.7%	11.3%	6.7%	7.9%	10.2%	3.4%	11.2%
	Places	8.9%	5.2%	15.7%	5.4%	5.6%	6.7%	7.9%	6.1%	13.8%	6.8%
	Restaurants	5.7%	7.8%	0.0%	8.6%	7.0%	13.3%	7.9%	2.0%	6.9%	6.3%
	Current Events	6.4%	2.6%	0.0%	8.6%	5.6%	0.0%	5.6%	10.2%	3.4%	5.4%
	Shopping	3.2%	7.8%	5.9%	2.2%	7.0%	6.7%	4.8%	4.1%	3.4%	4.9%

5.4 Group Interaction Behavior

5.4.1 Relationship Selection of Group Interaction

Mass online social networks are conducive to the spread of ideas and information, which has attracted the attention of scholars, advertisers and social activists. Existing research results focus on explaining the structure of online social networks, some scholars from the perspective

of the relationship between users, analyzed interactive behavior law. The strength of a relationship can make people think about how online social activities can be distributed on different types of connections, especially on different strength of connections. The arrangement of this section is as follows. Firstly, we introduce how to identify the relationship among users in online social networks, and then describe the relationship selection indexes based on the strength of the relationship between users, and finally analyze the actual case.

1. Relationship Identification in the Online Social Networks

Most of the traditional social networks are undirected. For example, in QQ or MSN, I am your friend, you are my friend. A relationship can only be established through mutual recognition. With the development of social media, the forms of social networks become more abundant. For example, Cameron Marlow et al.^[48] took Facebook as the research object and defined three types of connection based on a month's use record for users.

(1) Bidirectional connection: In the observation period, the user not only sends message to his/her friend, but also receives information from his/her friend.

(2) Unidirectional connection: A user only sends one or more messages to his/her friend without considering whether his/her friend replied to his message,.

(3) Maintain connection: If a user only cares about whether there is a friend on the other end without considering whether there is actual information exchange,. "Attention to the other side information" is indicated here either by Facebook's news alert service, or at least two times to visit his friend.

Huberman Bernardo et al^[49] also conducted a similar research on Twitter. Twitter has the characteristics of social network, and can distinguish the strong and weak relationship. Each user can specify a set of users he/she wants to follow (to see the information they send), or send messages directly to a particular person (the message is still open but labeled as sent to a particular person). The first type of interaction is a weak relationship, which causes the user to easily follow a lot of people without directly talking with them. While the second type of interaction corresponds to a strong relationship, in particular, for the user sending multiple messages directly to other user,.

2. Relationship Selection Indexes Under the Perspective of Strong and Weak Relationships

In graph theory, the definition of strong and weak relationships are as follows: strong

relationships refer to closely and frequent social contacts, and tend to be embedded in the dense areas of the network; weak relations refer to the casual and less social contact, and tend to cross the border of communities. At present, in online social network environment, relationship selection index based on the strong and weak relationships are mainly CN (common neighbors) index, AA (Adamic-Adar) index and so on.

CN index is the most simple similarity index based on local information, i.e. the possibility of the interaction between two nodes is larger if they have a lot of common friend nodes. CN index is defined as follows: v_x denotes a node in online social network, $\Gamma(x)$ denotes the set of v_x neighbor nodes. The possibility that two nodes v_x and v_y are connected is defined as the number of their common neighbor, namely

$$S_{xy} = |\Gamma(x) \cap \Gamma(y)| \quad (5-1)$$

Considering the impact of the node degree at both ends based on common neighbors, five possibility indexes that two nodes are connected are generated from different aspects.

- Salton index, also known as the cosine similarity:

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x k_y}} \quad (5-2)$$

wherein k_x or k_y respectively denotes the degree of node v_x or v_y .

- Jaccard index proposed by the Jaccard 100 years ago:

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (5-3)$$

- Sørensen index

$$S_{xy} = \frac{2 \times |\Gamma(x) \cap \Gamma(y)|}{|k_x + k_y|} \quad (5-4)$$

- Hub Promoted Index (HPI)

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\min\{k_x, k_y\}|} \quad (5-5)$$

Because the denominator is determined by the node that has smaller degree, hub nodes are easier to form adjacencies with other nodes.

- Hub Depressed Index (HDI)

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\max\{k_x, k_y\}|} \quad (5-6)$$

When measuring the possibility of a connection between two nodes by CN index, the degree of common neighbor node is ignored. It is not appropriate to calculate the relationship strength between the online social network nodes, which is easy to understand. For example, people who have many followers on micro blog are often stars or celebrities in one field, but if users choose such nodes to establish a connection, the effect is often not obvious. In contrast, if a node follows another user who has a small number of followers, while the third user also follows this user; in this case, the probability of establishing an interactive relationship between the second and third will be significantly improved. Based on this idea, Lada Adamic put forward AA index^[65]. AA index is defined as:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \quad (5-7)$$

The AA index assigns a weight value to each node according to the degree of its common neighbor node, which is equal to the reciprocal of the logarithm of the degree. Based on the AA index, inspired by the process of network resource allocation, Zhou Tao et al.^[53] proposed the RA index, which is defined as follows:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \quad (5-8)$$

The biggest difference between RA and AA index is how to assign the weight of the common neighbor nodes. The experimental results show that, when the average degree of the network is relatively small, the difference between RA and AA index is little, but when the average degree of the network is large, the RA index is superior to AA index.

The AA index and RA index respectively measure the weight of the common neighbor nodes by the reciprocal of the logarithm of its degree or the reciprocal of its degree. Because these indexes have lower complexity, it is available for large-scale online social networks. But for high precision request on small social network, RA index and AA index have some limitations. For this kind of problem, these indexes have local path index^[66], Katz index^[67], LHN-II index^[68], global random walk^[69], local random walk^[70], etc. It is worth noting that these algorithms have higher complexity and longer operation time despite their high accuracy, especially for global and local random walk. We don't introduce these indexes here in detail, interested readers can refer to the corresponding references.

3. Case Analysis

Cameron Marlow et al.^[48] researches group interaction on Facebook by the above

theory, according to the following steps:

- (1) Calculate the number of followers of an account;
- (2) Calculate the number of accounts interacting with an account.
- (3) Calculate all the accounts based on the above steps.

The study draws the following conclusions: First, on Facebook, even if the number of friends that a user claimed in his/her own data pages is large (about 500 people), but the total number of friends with actual contact is 10 to 40. Second, they believe that Facebook is able to promote the passive engagement, i.e. people keep in touch by reading news of his/her friends without communicating with each other. Huberman Bernardo et al. ^[49] came to the same conclusion on studying Twitter by the method similar to that of Cameron Marlow et al. As shown in Figure 5-23, even if the user has a large number of followers, but the number of friends is limited to 45. The establishment of friend relationship here is based on the relationship selection behavior on group interaction.

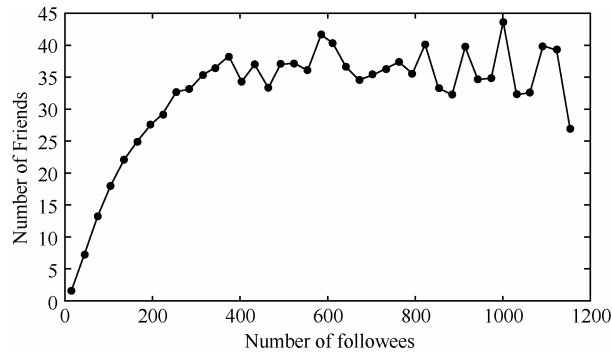


Figure 5-23 Relation graph of the number of strong relations and followers of online social network users

Cameron Marlow et al. ^[48] and Huberman Bernardo ^[49] et al. only conducted a qualitative research on the relationship selection of the group interaction. Ball Brian et al. ^[51] conducted the quantitative analysis and research for the relationship selection of the group interaction, using the reciprocity theory of social network. The results are shown in Figure 5-24.

Specific experimental steps are as follows:

- (1) Using LeaderRank algorithm to estimate the social status of each node in online social network;
- (2) Using the maximum likelihood model to calculate the connection probability of

any two nodes under a given social status gap.

As shown in the research, non-reciprocal edges are mostly from individuals with low social status to individuals with high social status, while reciprocal edges usually are generated between individuals with similar social status. Social status here refers to the status of node granted by certain sorting algorithms (e.g., Leader Rank) in given network environment, and high social status means more followers and larger influences.

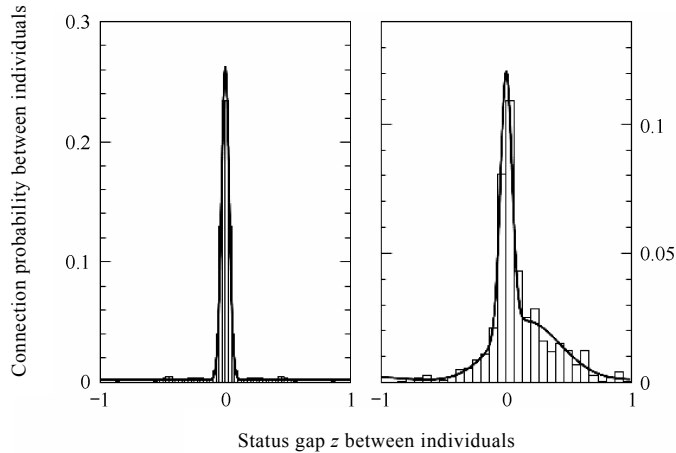


Figure 5-24 Relation graph of status gap between individuals and interaction establishment during group interaction

Figure 5-24(a) represents the relationship between the social status gap z of any two nodes and the connection probability from node v_x to node v_y . The probability of reciprocal connection presents a spike when $z=0$, which means that any two nodes with the same status are more likely to generate reciprocal connection. Figure 5-24(b) corresponds to the probability of non-reciprocal connections. It can be clearly seen that any two nodes with social status gap in the range of 0 to 1 are more likely to generate reciprocal connection, which means that unidirectional edge in online social network is usually from the nodes with low social status to the nodes with high social status. It is very prominent in microblogging social network platform in China. For example, we follow a celebrity in certain fields in microblog, but it's hard to be followed by the celebrity, making it impossible to create an interactive relationship. However, it is of high probability that celebrities follow each other, which shows that research conclusions of Ball Brian et al. are universal and applicable.

5.4.2 Content Selection of Group Interaction

The connection relationship between users in online social network is influenced not only by the relationship selection mechanism, but also by the content during interactive process. Crandall David et al. ^[54] researched the influence on similarity of two Wikipedia editors' behavior by their social behavior, and call the relationship selection mechanism as relationship influence and the content during interactive process as social influence. The results show that the similarity between content used by the two Wikipedia editors have significant differences before and after the interaction. Based on the data of Live Journal, a blog site, Backstrom Lars et al. ^[55] used friendship between individuals and network communities with obvious characteristics that individuals participated in to build bipartite graph. Based on membership closure theory in the triadic closure, they found that the number of individuals in a community (as the independent variable) and the probability that individuals tend to join a community (as the dependent variable), had a positive correlation. Brzozowski Michael et al. ^[56] applied the bipartite graph theory to Watercooler, a social network, and respectively built social relation network graph based on the behavior of group user interaction and community relationship graph based on individual user label information. They found that the recommendation effect based on community relationship diagram is worse than that of social relation network. In Watercooler, users can follow friends like Twitter and demonstrate themselves on a platform similar to Facebook. In a word, online social networks allow users to track or comment on some content, thus effectively focusing on relevant topics, and users can establish social relations corresponding to the content they follow or track. Therefore, the selection of group interaction is inseparable to social relations established in online social networks. The arrangement of this subsection is as follows. We first introduce the theoretical basis for researching the content selection of group interaction, and then the concrete experimental procedure, and finally case analysis.

Theoretical basis for the research of the content selection of group interaction is triadic closure theory and bipartite graph. The triadic closure theory solves the connection probability between two individuals having the same friend, while bipartite graph solves the probability that two individuals with similar social relationships are interested in a particular content. We need to apply the theory of community closure and membership closure based on the triadic closure theory. The theory of community closure solves the

connection probability between two individuals, which is the function of the number of communities they both participate in. For example, in microblog, user A and user B participate in the entertainment, food topic activities, and user A and user C participate in the entertainment topic activities; then how this difference affects the establishment of the relationship between individual users. The theory of membership closure puts forward another question from the opposite direction, i.e. how much is the probability that a person joins a particular community (as the function of the number of his/her friends who have participated in the community)? Also in microblog, user A has a friend who participates in the food topic activity, and user B has two friends who participate in the food topic activity, then how big the difference between probability that user A and user B are participating in the food topic activity.

The experimental procedures for researching the content selection in group interaction are as follows.

- (1) Select different network snapshot at equal time interval;
- (2) Build the bipartite graph for each network snapshot, and count the number of topics that each individual and two individuals participate in;
- (3) According to statistical data and taking time as the axis, analyze the influence of the community closure and membership closure on the user's relationship.

Case study of the content selection of group interaction: using triadic closure theory and membership closure theory and combining with the above research steps, Crandall David et al. ^[54] and Backstrom Lars et al. ^[55] respectively researched the Live Journal, a blog site, and Wikipedia. The results were shown in Figure 5-25.

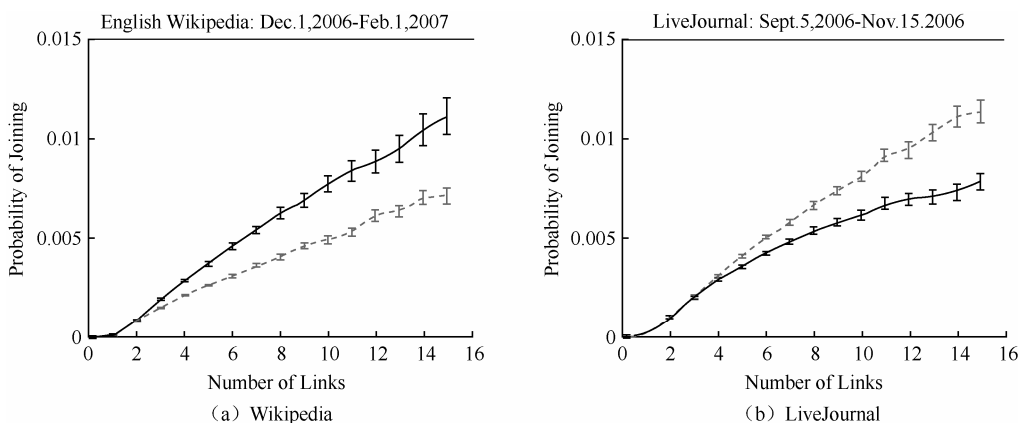


Figure 5-25 Member closure graph in online social network

Figure 5-25 (b) reflects the relationship between the number of a person's friends who have joined a community and the probability that such person joins this community. Figure 5-25 (a) is a similar analysis on Wikipedia. The nodes in bipartite graph respectively denote the user's account and user's statement homepage. If one editor leaves a comment on the homepage of another editor, an edge forms between the two editors. Each article of Wikipedia is a "community". Two experiments have come to the same conclusion, i.e. the probability that an individual joins the community increases with the increase of the number of common neighbors.

Using the same method on the Watercooler, a social network, Brzozowski Michael et al. analyzed the influence of group interaction content on the recommendation effect from the perspective of evaluating recommended schemes^[56]. There are three kinds of recommended schemes in the research:

- (1) Based on user behavior, i.e. individual A interacts with individual B frequently, indicating that the individual A is interested in B;
- (2) Based on network structure, including two research methods, i.e. collaborative percolation recommendation and recommendation based on the structural hole theory;
- (3) Based on user tag similarity, Watercooler allows users to label themselves with interest tags, which reflect the user's individual characteristics to some extent.

An interesting conclusion from the research is that the recommendation effect based on the user's tag similarity is not as good as that of network structure. The research team concluded that user's content label only indicates that he/she wants to become the person with such properties, but not "id". The network structure is the mapping of the real social structure as it forms by the interaction behavior between the real individuals. Therefore, compared with the former, it has better authenticity and thus better recommendation effect.

5.4.3 The Time Law of Group Interaction

In this subsection, we first introduce the time law of group interaction in online social networks, and then give two kinds of classical interpretation models for the time law.

The analyses of temporal characteristics of human behavior in online social networks mainly focus on the time interval distribution. For large-scale online social data sets, the dynamic mechanism of human behavior based on the time interval distribution of behavior is helpful to understand many socioeconomic phenomena, such as resource allocation, traffic control, epidemic forecasting as well as forecasting, emergency management,

personalized recommendation in economic activities, etc. With the development of information and network technology, social network tools (QQ, blog, forum, microblog, WeChat, etc.) are emerging, making it possible to conduct empirical research and modeling for the time interval distribution in large scale network for human behavior. Through researching on the time characteristics of online video on demand ^[57], online games ^[58], and posting micro blogs ^[59], it is found that significant difference exists between the time interval distribution and the negative exponential distribution in the traditional environment. The time interval distribution shows obvious power-law distribution characteristic, i.e. “the long tail” effect. The reason is that the behavior generation of the traditional environment research object satisfies Poisson flow, and behavior decision strategies is mainly first come first served (FCFS), such as research on traffic ^[60] and busy line in communication ^[61]. However, two general characteristics of the research object in the network environment make its behavior generation no longer satisfy the Poisson flow. Firstly, because of user interaction behavior in network environment, the network social relations formed thereby make the behavior generation in non-overlapping time domains no longer independent, such as online games. Secondly, users’ network behavior is embedded in real behavior. The behavior priority sequence based on user’s individual characteristics leads to paroxysmal characteristics of the user’s network behavior, i.e. the high-density outbreak and long-time waiting, such as information exchange on instant communication platform. Decisions are no longer made mainly by FCFS. Therefore, Barabasi Albert-Laszlo et al. ^[62] built the dynamic model of behavioral decision strategy, taking highest priority first (HPF) as the main and random selection as the secondary, well explained the power-law distribution mechanism of waiting time in variable task queue length situations. Vazquez Alexei et al. ^[63] further researched the characteristics of task waiting time and respectively put forward the two universal class views with power exponent of 1 and 1.5 for variable and immutable queue length situation. According to the change law over time of individuals’ interest in participating in activities, Han Xiao Pu et al. ^[64] put forward the model of human behavior time law in online social networks based on interest and motivation, which shows that the time law follows power law distribution.

1. Queuing Model Based on the Highest Priority

The theoretical hypothesis of Barabasi based on the highest priority model is as follows: the individual generates priority-based task queue sequence according to the importance of L tasks to be performed, and performs the task with the highest priority

within each time step. The task with the highest priority disappears from the queue thereafter, and a new task is added to the queue with priority x_i . Given the randomness of individuals' performance of tasks, Barabasi introduced variable γ ($0 \leq \gamma \leq \infty$) when creating the model. $\gamma = \infty$ indicates the user performs tasks exactly according to the priorities of tasks, and $\gamma = 0$ indicates the user performs tasks exactly by random selection. On the basis of this hypothesis, assuming the probability for performing a task within unit time is $\Pi(x) \sim x^{-\gamma}$, and the probability of performing the task with priority x at waiting time t is:

$$f(x, t) = (1 - \Pi(x))^{t-1} \Pi(x) \quad (5-9)$$

Thus, the average waiting time $\tau(x)$ for the task with priority x is:

$$\tau(x) = \sum_{t=1}^{\infty} t f(x, t) = \frac{1}{\Pi(x)} \sim \frac{1}{x^{-\gamma}} \quad (5-10)$$

$$P(\tau) \sim \frac{\rho(\tau^{-1/\gamma})}{\tau^{1+1/\gamma}} \quad (5-11)$$

wherein $P(\tau) \sim \tau^{-1}$ when $\gamma \rightarrow \infty$ and $P(\tau)$ converge in exponential distribution when $\gamma \rightarrow 0$.

2. Time Law Model of Human Online Activity Based on Interest-driven

The above models are the first quantitative analysis of the time distribution of the online social network activity, which inspired the enthusiasm of the follow-up researchers. On the basis of this, Han Xiao Pu et al. created a model of the time law of human online activity based on interest-driven. There are two points of the theory of the model, respectively:

(1) The probability of doing certain behavior changes after a person doing it. Consider the simplest case, i.e. assuming that the change rate is the same if the change trend is identical.

(2) There are two thresholds in the time interval of the event. Time interval being too small or too large will change the probability of executing behavior. That is, if the interval is too small, the probability of doing the behavior will be reduced; otherwise, the probability increases at the adverse proportion.

Based on this, the rules of the model are: the probability of occurrence of an action event is $r(t)$ at time t (time discreteness); $r(t)$ will have to update with each occurrence of the event according to update rules of $r(t+1) = a(t)r(t)$, wherein the value of $a(t)$ is as

follows:

$$a(t) = \begin{cases} a_0, & \tau_i \leq T_1 \\ a_0^{-1}, & \tau_i \geq T_2 \\ a(t-1), & T_1 < \tau_i < T_2 \end{cases} \quad (5-12)$$

wherein $T_1 \leq T_2$, $0 < a_0 < 1$.

According to the above rules, the situation that time interval between the adjacent two events is less than or equal to T_1 occurs mostly when $r(t)$ equals to or approaches T_1^{-1} . Therefore, the value of T_1 determines the maximum value of $r(t)$, the value of T_2 determines the minimum value of $r(t)$ in the vicinity of T_2^{-1} . If T_1 and the minimum time scale of the model are enlarged by the same ratio, and the ratio of T_1 and T_2 remains unchanged, then the average ratio between the maximum and minimum value of $r(t)$ is unchanged, thus the new time interval distribution is the same as that of the original, and T_1 represents the minimum effective time scale of the model. Therefore, in the following discussion, we assume $T_1 = 1$.

In numerical simulation, the initial value of $r(t)$ is fixed at 1.0. By means of numerical simulation, when the difference between T_1 and T_2 is over three magnitudes, it can generate the time interval distribution close to the power-law with the power exponent as -1. With the reduction of T_2 , the distribution gradually deviates from the power-law to the exponential distribution. When T_1 equals to T_2 , the larger a_0 also causes the distribution curve to deviate from the power-law, as shown in Figure 5-26.

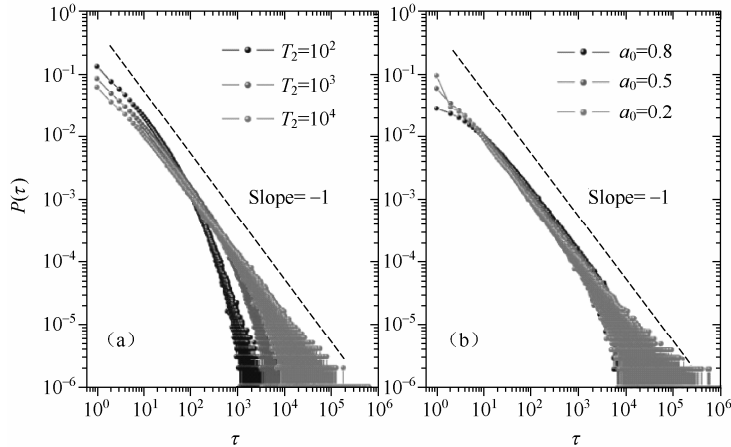


Figure 5-26 Time interval distribution generated by self-adaptive interest model

In this model, $r(t)$ is limited to a certain range and exhibits a quasi-periodic variation

because of the existence of two thresholds as shown in Figure 5-27. Only consider the stage of $r(t)$ reduction in a cycle, $r(t) = r_m a_0^i$ wherein r_m denotes the initial value (also the maximum value) of $r(t)$ and takes value around T_1^{-1} (but with mean smaller than T_1^{-1}) $i = 0, 1, 2, \dots$. The time interval distribution between two adjacent events is

$$P(\tau) = I^{-1} \sum_{i=0}^I (1 - r_m a_0^i)^{\tau-1} r_m a_0^i \quad (5-13)$$

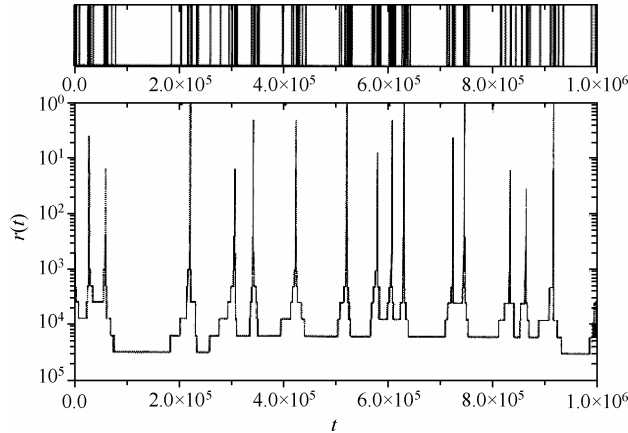


Figure 5-27 Pattern and interest change curve of events generated by self-adaptive interest model

The above theoretical model provides a solid theoretical basis for the interpretation of the power-law time distribution of human behavior in online social network environment.

5.5 Summary

As an important channel for social interaction in the Web 2.0 era, online social networks are used more and more widely in people's everyday life. With the online social networks, people can link with others, share information and happiness through the linkage, and interact with each other on the topics of common interest. Users' social networking behavior is the outward manifestation of their motivation. The characteristics and laws of users' behavior in social network can be used to analyze the internal mechanism of users' online social networking behavior, thereby provide knowledge for online social networking service provider to innovate their business model as well as theoretical foundations to monitor the online public sentiment.

Despite the fruitful theory and application results have been achieved by research on social network users' behavior, we think the following issues need to be further researched

and explored.

(1) For adoption and loyalty for social network, the existing research on the adoption and loyalty behavior of online social networks focuses on comprehensive social networking platforms (e.g., Facebook), with little results for the vertical and mobile social networking platforms. In addition, the technical properties (e.g., interactivity) of online social networks and the influence of users' psychological experience (e.g., social presence) on adoption and loyalty behavior are also challenging issues for the future study.

(2) For individual behavior of social network users, the existing research on the individual behavior in online social networks focuses on the individual behavior on social network platform like Facebook, such as the information posting behavior, information search behavior and information browsing behavior. Attracted by the rapid and efficient contact with target users by social network, in recent years, many enterprises advertise and provide customer services through the online social network platforms, and online social network platforms gradually develop service and business function based on information posting, entertainment, friend making, etc. The modeling of individual behavior involving social and business behavior needs to be further researched.

(3) For group behavior of social network users, the virtual communities in social network are usually formed by the strangers with no offline connection. The formation mechanisms of the virtual communities with distinct structures are still unclear. As the research on mutual influence between users in social network are usually conducted on the basis of network structure, the influence mechanism between users involving social properties needs to be further researched.

References

- [1] Davis Fred. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 1989, 13 (3): 319-340.
- [2] Kwon Ohbyung, Wen Yi Xing. An empirical study of the factors affecting social network service use[J]. *Computers in Human Behavior*, 2010 (26): 254-263.
- [3] Ernst Claus-Peter, Jella Pfeiffer, Franz Rothlauf.. The Influence of Perceived Belonging on Social Network Site Adoption[C]. *Proceedings of the Nineteenth Americas Conference on Information Systems (ACIS 2013)*, Chicago, Illinois, USA: 1-10.
- [4] Nikou Shahrokh, Bouwma Harry. Ubiquitous use of mobile social network services[J]. *Telematics and Informatics*, 2014 (31): 422-433.

- [5] Deb Sledgianowski, Songpol Kulviwat. Social Network Sites: Antecedents of User Adoption and Usage[C]. Proceedings of the 14th Americas Conference on Information Systems (AMCIS 2008), Toronto: Paper 83.
- [6] Dai Bao, Liu Yezheng. Research on the usage intention of SNS based on Technology Acceptance Model and perceived popularity [J]. Science & Technology Progress and Policy, 2012 (12): 47-51.
- [7] Icek Ajzen. The theory of planned behavior[J]. Organizational behavior and human decision processes, 1991, 50(2): 179-211.
- [8] Baker, Rosland White, Kathcrinc. Predicting adolescents' use of social networking sites from an extended theory of planned behaviour perspective[J]. Computers in Human Behavior, 2010 (26):1591-1597.
- [9] Emma Pelling, Katherine White. The theory of planned behavior applied to young people's use of social networking web sites[J]. Cyber Psychology & Behavior, 2009, 12(6): 755-759.
- [10] Chang Ya Ping, Zhu Dong Hong. Understanding social networking sites adoption in China: A comparison of pre-adoption and post-adoption[J]. Computers in Human Behavior, 2011 (27): 1840-1848.
- [11] Goh Say Leng, Suddin Lada, Mohd Zulkifli Muhammad, et al. An Exploration of Social Networking Sites (SNS) adoption in malaysia using technology acceptance Model (TAM), Theory of Planned Behavior (TPB) and Intrinsic Motivation [J]. Journal of Internet Banking and Commerce, 2011, 16 (2):1-27.
- [12] Anol Bhattacharjee. Understanding information systems continuance: an expectation-confirmation model[J]. MIS quarterly, 2001, 25 (3): 351-370.
- [13] Young Sik Kang, Soongeun Hong, Heeseok Lee. Exploring continued online service usage behavior: The roles of self-image congruity and regret[J]. Computers in Human Behavior, 2009 (25): 111-122.
- [14] Yin Guopeng, Yang Bo. Theoretical model and empirical research of SNS user continued behavior [J]. China Journal of Information Systems, 2010, 4 (1): 53-64.
- [15] Chen Yao, Shao Peiji. Empirical research on Social Web site continued usage: based on improved expectation confirmation model [J]. China Journal of Information Systems, 2011, 8 (1): 23-33.
- [16] Shin Soo, Hall Dianne. Identifying factors affecting SNS users as a temporary or persistent user: An empirical study[C]. Proceedings of the Seventeenth Americas Conference on Information Systems (AMCIS2011), Detroit, Michigan, USA: 316.
- [17] Li Qian, Hou Bimei. Research on the user continued usage intention of mobile social network based on DM and ECM-IT [J]. China Journal of Information Systems, 2013 (12): 50-59.
- [18] Mihaly Csikszentmihalyi. Beyond Boredom and Anxiety[M]. San Francisco: Jossey-Bass Publishers, 1975.

- [19] Zhou Tao, Li Hong Xiu, Liu Yong. The effect of flow experience on mobile SNS users' loyalty[J]. Industrial Management & Data Systems, 2010, 110 (6): 930-946.
- [20] Lin Hsu Chia, Chen Wu Cou. Understanding users' continuance of Facebook: An integrated model with the unified theory of adoption and usage of technology, expectation disconfirmation model, and flow theory[J]. International Journal of Virtual Communities and Social Networking, 2011, 3 (2): 1-16.
- [21] Chang Ya Ping, Zhu Dong Hong. The role of perceived social capital and flow experience in building users' continuous usage intention to social networking sites in China[J]. Computers in Human Behavior, 2012 (28): 995-1001.
- [22] Wu Yi, Wang Zheng, Chang Klarissa, Xu Yun Jie. Why People Stick to Play Social Network Site Based Entertainment Applications: Design Factors and Flow Theory Perspective[C]//Pacific Asia Conference on Information Systems (PACIS 2010), Taipei, Taiwan: 1041-1050.
- [23] Chang Chiao Chen. Examining users' intention to continue using social network games: A flow experience perspective[J]. Telematics and Informatics, 2013, 30 (4): 311-321.
- [24] Ryan Tracii, Xenos Sophia. Who uses Facebook? An investigation into the relationship between the BigFive, shyness, narcissism, loneliness, and Facebook usage[J]. Computers in Human Behavior, 2011, 27(5): 1658-1664.
- [25] Moore Kelly, McElroy James C. The influence of personality on Facebook usage, wall postings, and regret[J]. Computers in Human Behavior, 2012, 28 (1): 267-274.
- [26] Benevenuto Fabricio, Rodrigues Tiago, Cha Meeyoung, Almeida Virgilio et al. Characterizing user behavior in online social networks[C]. Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009: 49-62.
- [27] Golder Scott, Wilkinson Dennis, Huberman Bernardo. Rhythms of social interaction: Messaging within a massive online network[M]. Communities and Technologies 2007. Springer London, 2007: 41-66.
- [28] Maia Marcelo, Almeida Jussara, Almeida Viriglio. Identifying user behavior in online social networks[C]. Proceedings of the 1st workshop on Social network systems. ACM, 2008: 1-6.
- [29] Gyarmati Laszlo, Trinh Tuan Anh. Measuring user behavior in online social networks[J]. Network, IEEE, 2010, 24 (5): 26-31.
- [30] Mediabistro, October 18, 2011. [http://www.mediabistro.com/alltwitter/costolo-future-of-](http://www.mediabistro.com/alltwitter/costolo-future-of-twitter_b14936)
- [31] [twitter_b14936](http://blog.twitter.com/2012/03/twitter-turns-six.html). Source: <http://blog.twitter.com/2012/03/twitter-turns-six.html>, accessed October 9, 2012.
- [32] Shriver Scott K, Nair Harikesh S, Hofstetter Reto. Social ties and user-generated content: Evidence from an online social network[J]. Management Science, 2013, 59 (6): 1425-1443.

- [33] Toubia Olivier, Stephen Andrew T. Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter? [J]. *Marketing Science*, 2013, 32 (3): 368-392.
- [34] Wang Yi-Chia, Burke Moira, Kraut Robert. Gender, topic, and audience response: an analysis of user-generated content on facebook [C]. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013: 31-34.
- [35] Yarkoni Tal. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers[J]. *Journal of research in personality*, 2010, 44 (3): 363-373.
- [36] Qiu Lin, Lin Han, Ramsay Jonathan, Yang Fang. You are what you tweet: Personality expression and perception on twitter[J]. *Journal of Research in Personality*, 2012, 46 (6): 710-718.
- [37] Boyle Kris, Johnson Thomas. MySpace is your space? Examining self-presentation of MySpace users[J]. *Computers in Human Behavior*, 2010, 26 (6): 1392-1399.
- [38] OpenSource. (2010). OpenNLP: <http://opennlp.apache.org>.
- [39] David Blei Andrew Y. Ng, Michael I. Jordan. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 (Jan), 993-1022.
- [40] Pennebaker James, Francis, Booth. *Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program* [J]. Mahwah (NJ), 2001, 7.
- [41] Golbeck Jennifer, Cristina Robles, and Karen Turner. Predicting personality with social media[C]. *CHI' 11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2011: 253-262.
- [42] Nguyen Thin, Phung Dinh, Adams Brett, Venkatesh Svetha. Towards Discovery of Influence and Personality Traits through Social Link Prediction[C]. *ICWSM*. 2011.
- [43] Nowson Scott. *The Language of Weblogs: A study of genre and individual differences*[M]. Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, UK., 2006.
- [44] Morris Meredith Ringel, Teevan Jaime, Panovich Katrina. What do people ask their social networks, and why?: a survey study of status message q&a behavior[C]. *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2010: 1739-1748.
- [45] Van Zwol Roelof: Who is looking? [C]. *Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence*. IEEE Computer Society, 2007: 184-190.
- [46] Tifferet Sigal, Vilnai-Yavetz Iris. Gender differences in Facebook self-presentation: An international randomized study[J]. *Computers in Human Behavior*, 2014, 35: 388-399.
- [47] Morris Meredith Ringel, Teevan Jaime, Panovich Katrina. A Comparison of Information Seeking Using Search Engines and Social Networks[J]. *ICWSM*, 2010, 10: 23-26.
- [48] Cameron Marlow., Byron, L., Lento, T., and Rosenn, I.: 'Maintained relationships on Facebook', Retrieved February, 2009, 15: 2010.
- [49] Huberman Bernardo, Romero Daniel, Wu Fang.: 'Social networks that matter: Twitter under the

- microscope', arXiv preprint arXiv: 0812. 1045, 2008.
- [50] Burke Moria, Kraut Robert, Marlow Cameron.: 'Social capital on Facebook: Differentiating uses and users', in Editor (Ed.) ^ (Eds.): 'Book Social capital on Facebook: Differentiating uses and users' (ACM,2011, edn.) : 571-580.
 - [51] Ball Brian, Newman M E.: 'Friendship networks and social status', Network Science, 2013, 1 (1): 16-30.
 - [52] Zhang Jun Fu.: 'Tipping and Residential Segregation: A Unified Schelling Model*', Journal of Regional Science, 2011, 51 (1): 167-193.
 - [53] Zhou Tao, Lu Lin Yuan, ZhangYi Cheng Z.: 'Predicting missing links via local informat-ion', The European Physical Journal B, 2009, 71 (4): 623-630.
 - [54] Crandall David, Cosley Dan, Huttenlocher Daniel, Kleinberg Jon, Suri Siddharth.: 'Feedback effects between similarity and social influence in online communities', in Editor (Ed.) ^ (Eds.): 'Book Feedback effects between similarity and social influence in online communities' (ACM, 2008, edn.) : 160-168.
 - [55] Backstrom Lars, Huttenlocher Dan, Kleinberg Jon., Lan Xiangyang.: 'Group formation in large social networks: membership, growth, and evolution', in Editor (Ed.) ^ (Eds.): 'Book Group formation in large social networks: membership, growth,and evolution' (ACM, 2006, edn.): 44-54.
 - [56] Brzozowski Michael, Romero Daniel.: 'Who Should I Follow? Recommending People in Directed Social Networks', in Editor (Ed.) ^ (Eds.): 'Book Who Should I Follow? Recommending People in Directed Social Networks' (2011, edn.).
 - [57] Zhou Tao, Kiet H A T, Kim B J, Wang B H, Holme P.: 'Role of activity in human dynamics', EPL (Europhysics Letters), 2008, 82 (2): 28002.
 - [58] A Grabowski, N Kruszezwska, R A Kosiński Dynamic phenomena and human activity in an artificial society, Physical Review E, 2008, 78 (6): 066110.
 - [59] Yan Qiang, Wu Lian Ren, Zheng Lan.: 'Social network based microblog user behavior analysis', Physica A: Statistical Mechanics and its Applications, 2013, 392 (7): 1712-1723.
 - [60] Van Lint, Hans. Reliable travel time prediction for freeways. Netherlands TRAIL Research School, 2004.
 - [61] Beneš Vaclaw. E. Mathematical theory of connecting networks and telephone traffic. Academic Pr, 1965.
 - [62] Barabasi , Albert-Laszlo. 'The origin of bursts and heavy tails in human dynamics', Nature, 2005, 435(7039): 207-211.
 - [63] Vazquez Alexei, Oliveira Joao Gama, Dezso Zoltan, Goh Kwang-II, Kondor Imre, Barabasi Albert-Laszlo. : 'Modeling bursts and heavy tails in human dynamics', Physical Review E, 2006, 73

(3):036127.

- [64] Han Xiao Pu, Zhou Tao, Wang Bing Hong. Modeling human dynamics with adaptive interest[J]. New Journal of Physics, 2008, 10 (7): 073010.
- [65] Lada Adamic, Eytan Adar . Friends and neighbors on the web [J]. Social Networks, 2003, 25 (3): 211-230
- [66] Spring Ncil, Mahajan Ratul, Wetherall David, Anderson Thomas Measuring ISP topologies with rocketful [J]. IEEE/ACM Transactions on Networking, 2004, 12 (4): 2.
- [67] Leo Katz . A new status index derived from sociometric index[J]. Psychometrika, 1953, 18 (1): 39-43.
- [68] Leicht E A, Holme Petter, Newman M E J .Vertex similarity in networks[J]. Physical Review E, 2006,73 (2): 026120.
- [69] Douglas Klein, M. Randic. Resistance distance[J]. Journal of Mathematical Chemistry, 1993, 12 (1):81-95.
- [70] Liu Wei Ping, Lu Lin Yuan. Link prediction based on local random walk[J]. Europhysics Letters, 2010,89 (5): 58007.

Social Network Sentiment Analysis

6.1 Introduction

With the rapid development of Internet, network has become a main resource for users to obtain information and post opinions. Text can be divided into two types: objective description information mainly used to give objective description of events, products, etc., and subjective information mainly generated from user's comments on people, events, products, etc. Subjective information expresses user's emotional color and emotional orientation, such as "positive", "negative" or "neutral". Sentiment analysis (aka opinion mining) refers to the process of analyzing, processing and summarizing subjective information. It firstly originates from the natural language processing domain, and mainly researches and judges the sentiment orientation of text according to syntactic and semantic rule. With the rise and rapid development of social network, sentiment analysis gradually applies to many other research domains, such as text mining and web data mining, and extends to management science, social science and other disciplines. Now, it is widely used in product comments, public sentiment monitoring and information prediction.

This chapter will systematically introduce sentiment analysis techniques in social network. Section 6.2 introduces the sentiment analysis techniques for regular long text like news, reports and so on. Section 6.3 gives detailed introduction of main techniques for sentiment analysis in social network according to the short text characteristics in social network and the influence of social network link structure and group interaction on users' sentiment. Section 6.4 introduces the sentiment summary technique and sentiment analysis

technique based on transfer learning.

6.1.1 History of Sentiment Analysis

Sentiment analysis firstly originated from the analysis of words with emotional color. For example, “nice” usually is commendatory while “ugly” is derogatory. Although many researchers realized the importance of sentiment analysis, sentiment analysis techniques developed slowly before 1990s mainly due to lack of available data.

With the rise of Internet, network becomes the major source to obtain information. Massive available data like news and reports brings breakthrough for sentiment analysis technique. Professor Jance Wiebe from University of Pittsburgh firstly analyzed author’s subjective opinion in 1994 and divided sentences into objective and subjective text. Objective sentences are used to describe objective facts, while subjective sentences are used by authors to express their view, opinion, attitude and so on ^[1]. Sentiment analysis was firstly proposed by Sanjiv Das and Mike Chen ^[2] in their research on stock market text in 2001. They researched the messages in message board of stock market and defined sentiment as the positive and negative opinions in message. In 2003, Kushal Dave et al. ^[3] firstly used the word opinion mining, which aims to automatically extract the attributes of products (e.g., weight and feature) and to mine positive, neutral and negative opinion for each attribute. From then on, sentiment analysis and opinion mining were widely used in academic research. The main resources for sentiment analysis in that period were long text like news and reports whose grammar are regular enough for analysis and processing.

With the rise of Web 2.0 and social network, users can express their views and opinions at any moment, which contributes massive corpus for sentiment analysis as well as brings many new problems and challenges. Compared with long text like news and reports, text information in social networks has short length, irregular syntax rules, big data noise and, in particular, lots of popular Internet slang, making sentiment analysis more difficult. At the same time, group characteristics in social network as well as link and interaction characteristics among groups also brings a new research area for traditional sentiment analysis.

In 2008, Pang Bo and Lillian Lee regarded sentiment analysis and opinion mining as unified term in their summary ^[4] and defined them as extracting opinion, sentiment, and subjectivity in text. In 2012, Liu Bing defined sentiment analysis (aka opinion mining) as analyzing opinions, sentiments, attitudes, emotions and other subjective information

contained in user text related to products, services, events, topics and so on [5].

In this book, we also regard sentiment analysis and opinion mining as the same term.

6.1.2 Sentiment Definition and Classification

According to the text granularity, sentiment analysis can be divided into three levels: article level, sentence level and word level. Article level sentiment analysis regard the whole article as the target for sentiment analysis to mine its sentiment orientation on an event or a product, generally expressed by ternary classification (positive, neutral and negative) or numerical evaluation (e.g., 1~5). Sentiment analysis at article level aims to mine the author's overall attitude, but neglect sentiment polarity in some sentences. For example, in a sentiment analysis for a report on iPhone, it is difficult to mine some negative comments despite the overall positive attitude of the author. Sentiment analysis at sentence level regards sentence as the independent target for sentiment analysis, which can effectively cover the above shortage. For the sentence to be analyzed, firstly classify it as an objective sentence or subjective sentence, and then determine the sentiment polarity of subjective sentence. Sentiment analysis at word level aims to determine the sentiment polarity for each word, which is mainly used in constructing sentiment dictionary. However, it neglects the influence of context, thus is not available for differentiating the sentiment polarity of the same word in different context. For example, the word "high" in sentence "the price of cannon is high" is a negative word, but is positive in sentence "the iPhone's cost performance is high".

We define views and opinions to be extracted in sentiment analysis according to Reference [5].

Definition 6-1 (Opinion) is generally defined as a tetrad $\langle g, s, h, t \rangle$, wherein g denotes the sentiment object or target, s denotes the sentiment orientation, h denotes the opinion holder, t denotes time.

Sentiment orientation s (aka sentiment polarity) usually expressed by ternary classification: positive, neutral and negative.

We take a comment for iPhone in Sina Weibo as an example.

Example 6-1

User A: ①I bought an iPhone 5 yesterday. ②It has beautiful and fashionable appearance, high pixel camera, good photograph quality. I like it very much. ③However, my classmate B thinks it is too expensive despite its pretty appearance. The cost

performance is low. — 2013-12-20.

In Example 6-1, sentence ① is an objective sentence describing the fact that the author bought a cellophane without any sentiment orientation. Sentence ② is a positive comment for iPhone with opinion holder as user A. Sentence ③ is a negative comment, but the opinion holder is user B. The post time of comment represents the time that user holds such opinion. Therefore, we can extract the following two opinions.

Example 6-2 Extract opinions in product comments in Example 6-1:

Opinion 1: $\langle g, s, h, t \rangle = \langle \text{iPhone}, \text{positive}, \text{user A}, 2013-12-20 \rangle$

Opinion 2: $\langle g, s, h, t \rangle = \langle \text{iPhone}, \text{negative}, \text{user B}, 2013-12-20 \rangle$

The task of sentiment analysis is to extract the opinions from comments. For example, extraction of Opinion 1 and Opinion 2 from comments in Example 6-1 involves comment target, sentiment orientation, opinion holder and comment time.

We can see from Example 6-1 that, although user A and user B comment the same product iPhone, user A concerns more about the appearance and camera performance while user B concerns more about the price. Therefore, they hold different sentiment tendencies. In 2004, Liu Bing introduced the concept of entity into sentiment analysis. Each entity contains many features or aspects to represent different attributes of product. Take Example 6-1 for example, regard iPhone as an entity with such attributes as appearance, quality, weight, battery and price. After introduction of entity, the target for sentiment analysis to be extracted is no longer the entity, but different attributes of entity. Therefore, we can extend the Definition 6-1 as follows.

Definition 6-2 (Entity-based opinion) An opinion about the entity is a quintuple $\langle e, a, s, h, t \rangle$ wherein e denotes the entity, a denotes different attributes of the entity, s denotes the sentiment orientation, h denotes the opinion holder, and t denotes time.

Entity-based sentiment analysis requires not only extracting the sentiment orientation of each entity, but also conducting sentiment analysis on each attribute. Opinions in Example 6-1 can extend as follows.

Example 6-3 Extract opinion in Example 6-1 based on entity.

Opinion 1: $\langle e, a, s, h, t \rangle = \langle \text{iPhone}, \text{appearance}, \text{positive}, \text{user A}, 2013-12-20 \rangle$

Opinion 2: $\langle e, a, s, h, t \rangle = \langle \text{iPhone}, \text{camera}, \text{positive}, \text{user A}, 2013-12-20 \rangle$

Opinion 3: $\langle e, a, s, h, t \rangle = \langle \text{iPhone}, \text{price}, \text{negative}, \text{user B}, 2013-12-20 \rangle$

Opinion 4: $\langle e, a, s, h, t \rangle = \langle \text{iPhone}, \text{cost performance}, \text{negative}, \text{user B}, 2013-12-20 \rangle$

Given a document set D , sentiment analysis contains the following tasks.

Task 1: (Entity extraction and classification) Extract all entities in D, and classify or group them. Each category represents an unique entity, such as entity iPhone.

Task 2: (Attribute extraction and classification) Extract attributes of each entity, and classify or group them. Each category represents an unique attribute of entity, such as the appearance, quality and camera of iPhone.

Task 3: (Opinion holder extraction) Extract the opinion holder of each opinion in Definition 6-2.

Task 4: (Time extraction) Extract the comment time for each opinion or the time that opinion holder expresses such opinion, and standardize the time. For example, relative time like “Yesterday” should transfer to GMT format.

Task 5: (Sentiment orientation extraction) Extract the sentiment orientation for evaluation object of opinion holder, expressed by ternary classification (positive, neutral and negative) or numerical evaluation (e.g., 1 ~ 5).

Task 5 is the main task of text sentiment analysis. In this chapter, we mainly introduce techniques for sentiment orientation extraction. Table 6-1 shows the common terms and relevant interpretations in this chapter without causing conflict.

Table 6-1 Common terms and relevant interpretations in this chapter

Term	Similar term or interpretations
Sentiment analysis	aka opinion mining
Sentiment orientation	aka sentiment polarity
Positive	means the sentiment orientation is positive
Negative	means the sentiment orientation is negative
Evaluation word	aka sentiment word, used to express author’s sentiment, such as “beautiful” and “fashionable”

6.1.3 Application of Sentiment Analysis

Sentiment analysis is widely used in domains of product comment, public sentiment monitoring and information prediction. Before users buy a product, they tend to check comments related to such product and make final decision by comparison with other products. As users do not have enough time and energy to browse all comments, various systems are developed for providing comments of all attributes and facilitating users to make final decisions by statistics, conclusion and inference. For example, Liu Bing et al. developed the OpinionObserver system to process product comments of online customers [6], and conduct visualized comprehensive quality comparison on several kinds of products. Wilson Theresa et al. developed the OpinionFinder system to automatically

identify subjective sentences and extract sentiment information therein^[7].

Public sentiment monitoring is another important application domain for sentiment analysis. As the social network has openness, virtuality, divergency, permeability, randomness and other characteristics, more and more users are willing to express their attitudes, which makes social network the main resource for generation and propagation of public sentiment. As everyone has speaking right in network, various topics and views related to national economy and people's livelihood can be posted at any time and forthwith spread in a "fission" manner. Opinion leaders can mobilize people with the same views, sentiments and appeals and rapidly mobilize the masses to participate in social activities offline, thereby forming social mobilization force. The integration of and interaction between virtual social network and real society have bigger and bigger direct influence on society. Therefore, the perception and analysis of people's sentiment and attitude in network are of great importance in maintaining national security and promoting social development.

Sentiment analysis plays an important role in information prediction. People's thoughts and actions are largely influenced by the occurrence of a new event or heated discussion over an event. Sentiment analysis technique can predict the development orientation of future events by analyzing and processing users' sentiment orientation in text, and has played an important role in economic and political domains. For example, Devitt Ann et al. predicted future financial orientation by performing sentiment polarity recognition on financial comment text^[8]. Kim Soo-Min et al. successfully predicted the results of U.S. president election in 2008 by analyzing a large number of corresponding network news comment^[9].

6.2 Sentiment Analysis Techniques

Sentiment analysis techniques in online social network mostly derives from text sentiment analysis. The text here, different from social network like Sina Weibo, usually refers to "long text", such as network news, netizen blogs and forum post. The research method for text sentiment analysis can be classified into semantic rule-based techniques, supervised learning-based techniques and topic model-based sentiment techniques, which will be introduced in this section in detail.

6.2.1 Semantic Rule-based Sentiment Analysis

From the perspective of part of speech, noun usually represents entity or its attribute

while adjective and adverb are usually used for expressing sentiment views, i.e. comment words usually consist of adjective and adverb. It is easy to calculate sentiment orientation of the author if sentiment polarity of all comment words is obtained. For example, considering the comment “iPhone has beautiful, new-fashioned, characteristic and fashionable appearance”, it is easy to determine author’s sentiment orientation as positive because sentiment words “beautiful, new-fashioned, characteristic and fashionable” used for expressing the author’s view are positive. This method relies on the sentiment dictionary labeled with sentiment orientation and is called as sentiment dictionary-based method.

Sentiment dictionary-based method can give a preliminary judgment of sentiment orientation of text, but is not applicable for all circumstances as no single sentiment dictionary can include all comment word, and some sentiment words have different polarity in different context. The semantic rule-based method can achieve sentiment classification by calculating the distance between comment words and seed words in sentiment dictionary (words indicating degree of sentiment orientation labeled in sentiment dictionary).

Essentially, semantic rule-based sentiment analysis techniques are unsupervised learning methods. Peter D.Turney proposed a typical algorithm SO-PMI^[10] in 2002. Only “excellent” and “poor” were used as the benchmark words for positive and negative comment, and sentiment classification was achieved by calculating the distance between comment words and the above two reference words based on pointwise mutual information.

Algorithm 6-1 SO-PMI algorithm
Input: Document d to be analyzed Output: Sentiment orientation of document d
<p>Steps are as follows:</p> <p>Step1. Extract comment word set W from the document to be analyzed by the means of labeling part of speech with mainly adjective and adverb as comment words. Note: comment words are generally used to express user’s orientation, but their orientation varies under different context.</p> <p>Step2. Select “excellent” and “poor” as the benchmark words. For each sentiment word $w_i \in W$, calculate its semantic orientation based on pointwise mutual information as follows:</p> $SO(w_i) = PMI(w_i, \text{excellent}) - PMI(w_i, \text{poor})$

Pointwise mutual information

$$\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1 \& \text{word}_2)}{P(\text{word}_1)P(\text{word}_2)}$$

wherein $P(\text{word}_1 \& \text{word}_2)$ denotes the probability that word_1 and word_2 appear at the same time, $P(\text{word}_i)$ denotes the probability that word_i appears alone.

Step3. Calculate the average semantic orientation of sentence:

$$\text{SO}(W) = \frac{1}{|W|} \sum_{w_i \in W} \text{SO}(w_i)$$

If $\text{SO}(W) > 0$, the sentiment orientation is positive; otherwise, if $\text{SO}(W) < 0$, the sentiment orientation is negative.

SO-PMI algorithm uses pointwise mutual information to measure the distance between comment words and seed words. The basic principle is that bigger PMI value between comment words and seed words brings bigger probability that comment words and seed words appear at the same time, indicating more similar sentiment orientation.

We take Example 6-1 to show the text sentiment analysis steps of SO-PMI algorithm.

Example 6-4 Calculate sentiment orientation use example 6-1.

(1) First, label the comment by part of speech and extract the comment words as shown in Table 6-2.

Table 6-2 Comment words

No.	Comment words
Sentence 2	Fashionable, newfashioned, high pixel, good quality
Sentence 3	Beautiful, high press, low cost performance

(2) Second, calculate the pointwise mutual information between each comment word and reference word (i.e. “excellent” and “poor” by the formula $\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1 \& \text{word}_2)}{P(\text{word}_1)P(\text{word}_2)}$. Assume that the times of comment words and reference words in 1,000 corpora are as shown in table 6-3.

Table 6-3 The times of comment words and reference words

	excellent	poor	Times alone
Fashionable	57	3	97
Newfashioned	37	1	122
High pixel	23	2	85

(To be continued)

Continued table

	excellent	poor	Times alone
Good quality	28	4	76
Beautiful	46	3	97
Expensive	6	29	68
Low cost performance	1	52	79
Times alone	137	102	

The cross terms in Table 6-3 means the times that two words appear and the last column and row mean the time that a single word appears. For example, in 1,000 comments, the word “fashionable” appears 57 times with word “excellent”, but only 3 times with “poor”. Meanwhile, “fashionable” appears 97 times alone, “excellent” appears 137 times alone and “poor” appears 102 times alone. Therefore,

$$\text{PMI}(\text{“fashionable”, excellent}) = \log_2 \frac{P(\text{fashionable, excellent})}{P(\text{fashionable})P(\text{excellent})} = \log_2 \frac{0.057}{0.097 \times 0.137} = 2.1007$$

$$\text{PMI}(\text{“fashionable”, poor}) = \log_2 \frac{P(\text{fashionable, poor})}{P(\text{fashionable})P(\text{poor})} = \log_2 \frac{0.003}{0.097 \times 0.102} = -1.7216$$

Therefore,.

$$\text{SO}(\text{fashionable}) = \text{PMI}(\text{fashionable, excellent}) - \text{PMI}(\text{fashionable, poor}) = 3.8223$$

In a similar way, we can calculate the SO-PMI value of each comment word as shown in Table 6-4.

Table 6-4 Calculate SO-PMI values

Comment word w	$\text{PMI}(w, \text{excellent})$	$\text{PMI}(w, \text{poor})$	$\text{SO}(w)$
Fashionable	2.1007	-1.7216	3.8223
Newfashioned	1.1465	-3.6374	4.7839
High pixel	0.9819	-2.116	3.0979
Good quality	1.4272	-0.9546	2.3818
Beautiful	1.7914	-1.7216	3.513
High price	-0.6347	2.0639	-2.6986
Low cost performance	-3.436	2.69	-6.126

(3) Calculate overall SO value. According to the length of text, we can get the sentiment orientation for the document, Sentence 2 and Sentence 3.

- ① Sentiment orientation of the comment:

$$SO(W) = \frac{1}{|W|} \sum_{w_i \in W} SO(w_i) = 1.2535 > 0$$

Therefore, the overall polarity of comment is positive.

- ② Sentiment orientation of Sentence 2:

$$SO(W_2) = \frac{1}{|W_2|} \sum_{w_i \in W_2} SO(w_i) = 3.5215 > 0$$

Therefore, Sentence 2 is positive.

- ③ Sentiment orientation of Sentence 3:

$$SO(W_3) = \frac{1}{|W_3|} \sum_{w_i \in W_3} SO(w_i) = -1.7705 < 0$$

Therefore, Sentence 3 is negative.

SO-PMI was the first model analyzing text sentiment by unsupervised learning algorithm and applied it to the domains of car comments and movie comments. The average precision in car comments is 84%, and 66% in movie area. Many research were performed based on it thereafter. For example, Ding Xiaowen et al. proposed a whole dictionary-based opinion mining method^[11] in 2008. For each sentence s , they thought it contained multiple features and corresponding sentiment words. The orientation of each feature can be determined according to the following formula.

$$\text{score}(f) = \sum_{w_i: w_i \in s \cap w_i \in V} \frac{w_i \cdot \text{SO}}{\text{dis}(w_i, f)}$$

wherein w_i denotes the sentiment word, V denotes the set of sentiment word, $\text{dis}(w_i, f)$ denotes the distance between sentiment word w_i and feature f , $w_i \cdot \text{SO}$ denotes the sentiment polarity of w_i with positive as 1 and negative as -1. For each feature f , $\text{score}(f) > 0$ means the sentiment polarity for feature f is positive, while $\text{score}(f) < 0$ means negative, otherwise neutral. Meanwhile, they considered the influence of negative and adversative words on improving precision. For example, “negative word + negative = positive”, “negative + positive = negative”. Adversative words, such as “but”, “whereas”, always imply an opposite sentiment orientation.

In 2005, Kim Soo-Min and Eduard Hovy proposed an method to collect sentiment words based on WordNet semantic distance^[12]. They firstly collected 34 adjectives and 44 adverbs as seed words, and used WordNet to extend sentiment words. The basic idea is that

the synonyms and antonyms of a sentiment word are sentiment word too. For each word w , they used the following formula to decide the its sentiment polarity.

$$\arg \max_c P(c | w) \cong \arg \max_c P(c | \text{syn}_1, \text{syn}_2 \cdots, \text{syn}_n)$$

wherein c denotes the target classification (sentiment word or non-sentiment word), w denotes the target word, syn_i denotes the synonym or antonym of w in WordNet. According to Bayesian formula, we have

$$\begin{aligned} \arg \max_c P(c | w) &= \arg \max_c P(c)P(w | c) \\ &= \arg \max_c P(c)P(\text{syn}_1, \text{syn}_2, \text{syn}_3, \cdots, \text{syn}_n | c) \\ &= \arg \max_c P(c) \prod_{k=1}^m P(f_k | c)^{\text{count}(f_k, \text{synset}(w))} \end{aligned}$$

wherein f_k denotes the k feature of category c and belongs to synonym set of w ; $\text{count}(f_k, \text{synset}(w))$ denotes the number that f_k appears in synonym set of w . The sentiment orientation of w is decided by the results of classification.

6.2.2 Supervised Learning-based Sentiment Analysis

Supervised learning-based sentiment analysis firstly labels text polarity manually and use it as training dataset, and construct classifier based on machine learning technique, so as to perform sentiment classification for target text.

Pang Bo et al. firstly introduced machine learning techniques into text sentiment analysis on movie comment data in 2002. They firstly labeled 752 negative comments and 1,301 positive comments as training dataset, and use Naïve Byes, maximum entropy and support vector machines to perform sentiment classification for target text. The results showed that machine learning techniques can effectively improve the precision of sentiment analysis.

1. Naïve Byes Classification

Let $d = \{f_1, f_2, \cdots, f_n\}$ be a document, wherein f_i denotes the feature or attribute of document. Let c be the sentiment orientation of document where 1 means positive and -1 means negative. Given a target document d , Naïve Byes method firstly calculates the posterior probability using training data set, and use the maximum probability of c as the document orientation of d . That is

$$c = \arg \max_c P(c | d)$$

According to Naïve Bayesian conditional probability equation,

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}.$$

wherein $P(d)$ is the probability of document d . As $P(d)$ are independent of sentiment orientation c , $P(d)$ has no influence on the classification results. Assume f_i , the features of d , is independent of each other, then conditional probability

$$P(d|c) = P(f_1, f_2 \cdots f_n | c) = \prod_i P(f_i | c)$$

The Naïve Byes method firstly calculates the prior probability distribution $P(c)$ and the conditional probability $P(d|c)$ to obtain the sentiment classification results. The algorithm is as follows.

Algorithm 6-2 Sentiment analysis based on Naïve Bayesian method.
Input: Document dataset $D = \{d_i, c_i\}$ labeled with sentiment classification, wherein sentiment orientation of d_i is c_i with values as 1 or -1. Output: Sentiment orientation of target document d .
Steps of the algorithm are as follows. Step1. Calculate prior probability $P(c)$, and the conditional probability $P(f_i c)$: $P(c) = \frac{\#c}{N}$ wherein $\#c$ denotes the number of documents with sentiment orientation of c in D , N denotes the number of total documents. $P(f_i c) = \frac{P(f_i, c)}{P(c)}$ Step2. Calculate posterior probability: $P(c d) \propto P(c)P(d c) = P(c) \prod_i P(f_i c)$ Step3. Choose the maximized posterior probability as the output: $c = \arg \max_c P(c d)$

We use Example 6-5 to show how Naïve Bayesian model works. In this example, the training dataset contains 15 documents, and each document contains two features f_1 and f_2 . The feature f_i can denote the comment words in document or any other attributes showing user's sentiment orientation, and value $\{0,1,2\}$ denotes the times that f_i

appears in the document. c denotes the sentiment orientation with value as $\{-1, 1\}$, where 1 denotes positive and -1 denotes negative. We only use this example to show the steps for classifying user's sentiment by Naïve Bayesian algorithm. In real corpus, the number of training dataset and features are much bigger.

Example 6-5 Using the training dataset in Table 6-5 to calculate the sentiment orientation of $d = \{1, 0\}$.

Table 6-5 Training Dataset

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
f_1	0	0	0	0	0	1	1	1	1	1	2	2	2	2	2
f_2	0	1	1	0	0	0	1	1	2	2	2	1	1	2	2
c	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

Solution:

Step 1. Calculate the prior probability and conditional probability as follows:

$$P(c = 1) = \frac{9}{15}, \quad P(c = -1) = \frac{6}{15}$$

$$P(f_1 = 0 | c = 1) = \frac{2}{9}, P(f_1 = 1 | c = 1) = \frac{3}{9}, P(f_1 = 2 | c = 1) = \frac{4}{9}$$

$$P(f_2 = 0 | c = 1) = \frac{1}{9}, P(f_2 = 1 | c = 1) = \frac{4}{9}, P(f_2 = 2 | c = 1) = \frac{4}{9}$$

$$P(f_1 = 0 | c = -1) = \frac{3}{6}, P(f_1 = 1 | c = -1) = \frac{2}{6}, P(f_1 = 2 | c = -1) = \frac{1}{6}$$

$$P(f_2 = 0 | c = -1) = \frac{3}{6}, P(f_2 = 1 | c = -1) = \frac{2}{6}, P(f_2 = 2 | c = -1) = \frac{1}{6}$$

Step 2. Calculate the posterior probability:

$$P(c = 1 / d) = P(c = 1)P(f_1 = 1 | c = 1)P(f_2 = 0 | c = 1) = \frac{9}{15} \times \frac{3}{9} \times \frac{1}{9} = \frac{1}{45}$$

$$P(c = -1 / d) = P(c = -1)P(f_1 = 1 | c = -1)P(f_2 = 0 | c = -1) = \frac{6}{15} \times \frac{2}{6} \times \frac{3}{6} = \frac{1}{15}$$

Step 3. Select the maximized probability as output:

$$P(c = -1 / d) > P(c = 1 / d)$$

As the sentiment classification of document d is $c=-1$, document d is negative.

2. Maximum Entropy Model

The maximum entropy theory was firstly proposed by Edwin T. Jaynes in 1957^[14]. Its main idea is if we only know part of the whole information, we should choose the probability distribution that meets those conditions but has the maximum entropy. Assume $P(X)$ is the probability distribution of variable X , then its information entropy is

$$H(P) = -\sum_x P(x) \log P(x)$$

We can prove that the information entropy meets the following inequation

$$0 \leq H(P) \leq \log_2 N$$

wherein N denotes the number of possible values of X . If and only if X meets uniform distribution, the right inequation becomes equation, i.e. the entropy has the maximum value when X meets uniform distribution.

Given some constraints of X , there are many probability distribution meeting those constraints. Each probability distribution can be regarded as a model. The maximum entropy theory is to choose the model with the maximum entropy as the output model.

Example 6-6 Calculate information entropy.

Assume the values of variable X are in $\{a, b, c, d, e, f, g, h\}$.

(1) If other information is unknown, we only have the following constraint:

$$p(a) + p(b) + p(c) + p(d) + p(e) + p(f) + p(g) + p(h) = 1$$

There are many probability distribution meeting the above constraint, and each distribution has its information entropy, like the following two probability distribution:

① If X meets uniform distribution, i.e.

$$p(a) = p(b) = p(c) = p(d) = p(e) = p(f) = p(g) = p(h) = \frac{1}{8},$$

Then the information entropy is

$$H(P_1) = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3$$

② If X meets probability distribution $\left\{ \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64} \right\}$, the information entropy is

$$H(P_2) = -\left(\frac{1}{2}\log_2 \frac{1}{2} + \frac{1}{4}\log_2 \frac{1}{4} + \frac{1}{8}\log_2 \frac{1}{8} + \frac{1}{16}\log_2 \frac{1}{16} + \frac{1}{64}\log_2 \frac{1}{64} + \frac{1}{64}\log_2 \frac{1}{64}\right) = 2$$

and we have $H(P_1) > H(P_2)$.

(2) Based on constraint 1, we add another constraint:

$$p(a) + p(b) = \frac{1}{2}$$

There are also many distributions meeting the above two constraints. When X has uniform distribution on the premise of meeting these constraints, i.e.

$P(X) = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}\right\}$, the information entropy has the maximum value as follows

$$H(P_3) = -\left(2 \times \frac{1}{4}\log_2 \frac{1}{4} + 6 \times \frac{1}{12}\log_2 \frac{1}{12}\right) = 2.79$$

The information entropy of other distributions are all smaller than this distribution.

Given a training document D labeled with text polarity, we can extract a feature function from the training document as the constraints. Assume

$$F_i(d, c) = \begin{cases} 1, & \text{if } d \text{ and } c \text{ meets fact} \\ 0, & \text{otherwise} \end{cases}$$

then maximum entropy model can transfer to an optimization problem, i.e.

$$\max H(p(c|d)) = \max - \sum_{c,d} \tilde{p}(d) p(c|d) \log p(c|d) = \min \sum_{c,d} p(c,d) \log p(c|d)$$

then

$$\begin{cases} \sum_{c,d} p(c|d) \tilde{p}(d) F_i(d, c) = \sum_{c,d} \tilde{p}(c|d) \tilde{p}(d) F_i(d, c) \\ \sum_c p(c|d) = 1 \end{cases}$$

wherein $\tilde{p}(d)$ denotes the probability that document d appears in training dataset D , $\tilde{p}(c|d)$ denotes the probability that document d are classified into sentiment category c .

We give the solution for the above optimization problem, i.e. the format of maximum entropy model as follows. Please refer to Reference^[15] for detailed inference process.

$$P_{ME}(c|d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d, c)\right)$$

wherein

$$Z(d) = \sum_c \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d, c)\right)$$

and $\lambda_{i,c}$ denotes model parameter. Iteration-based numerical methods are effective ways to solve the maximized model parameters. Two typical methods are GIS (Generalized Iterative Scaling) algorithm and IIS (Improved Iterative Scaling) algorithm.

3. Support Vector Machines Method

Support vector machines (SVM) is a binary classification method with the basic model as defining the linear classifier with maximized interval in eigen space. Given eigen space, if we can find a hyperplane that can divide the entities in eigen space into different categories, the space is linear separable. The hyperplane have the following format: $w \cdot x + b = 0$ wherein w denotes the normal vector and b denotes the intercept. For example, Figure 6-1 shows a linear separable space where there is a hyperplane that can divide positive and negative data into two groups.

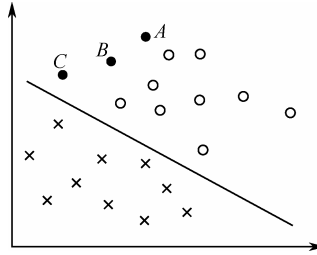


Figure 6-1 Linear separable space

Apparently, the hyperplane is not unique, and SVM tries to achieve classification by solving the optimal hyperplane with maximized interval. This problem can transfer to the following optimization problem:

$$\min \frac{1}{2} \|w\|^2$$

then for all training dataset:

$$y_i (w \cdot x_i) + b - 1 \geq 0$$

The steps for text sentiment classification of SVM algorithm is as follows.

Algorithm 6-3 Support vector machines model

Input: Document dataset labeled with sentiment category $D = \{d_i, c_i\}$ wherein c_i denotes the sentiment orientation of document d_i with values as 1 or -1.

Output: Hyperplane with maximized interval and classification decision-making function.

Step 1. Construct and solve an optimization problem:

$$\min \frac{1}{2} \|w\|^2$$

then

$$c_i(w \cdot d_i) + b - 1 \geq 0$$

Obtain optimal solution w^* and b^* .

Step 2. Construct a hyperplane

$$w^* \cdot x + b^* = 0$$

and a classification decision-making function

$$f(d) = \text{sign}(w^* \cdot x + b^*)$$

Therefore, the sentiment orientation of a given document d can be determined by classification decision-making function $f(d)$. If $f(d)=1$, document d is positive; if $f(d)=-1$, document d is negative.

6.2.3 Topic Model-based Sentiment Analysis

Many scholars introduce the emerging topic model into sentiment analysis to analyze users' sentiment or attitude towards certain topic or event. Topic model-based models like Probabilistic Latent Semantic Analysis (PLSA) ^[16] and Latent Dirichlet allocation (LDA) ^[17] add sentiment word variable based on topic model to identify the topic of document and author's sentiment orientation at the same time. PLSA model and LDA model are all Bayesian generative models. Please refer to "Pattern recognition and machine learning" for further understanding.

In 2010, Zhao Wayne Xin et al. proposed a maximum entropy LDA model to perform sentiment analysis ^[19]. They improved LDA model to identify the comment target and word. The generative model is as shown in Figure 6-2.

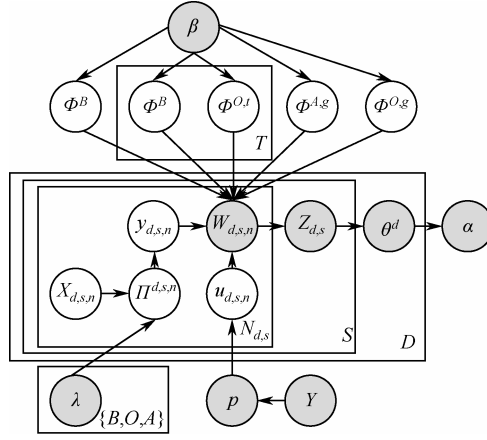


Figure 6-2 Maximum entropy LDA model

In this model, parameter α denotes the Dirichlet prior distribution for documents over topics, and β for topics over words. For each document d , its topic distribution follows a Dirichlet distribution with parameter α , i.e. $\theta^d \sim \text{Dir}(\alpha)$. For each topic, its distribution is the same as that in LDA model, which meets $z_{d,s} \sim \text{Multi}(\theta^d)$ wherein $\text{Multi}(\cdot)$ means a multinomial distribution. The maximum entropy LDA model, according to parameter β , generates four different target distribution: background model ϕ^B , general aspect model $\phi^{A,g}$, topic aspect model $\{\phi^{A,t}\}_{t=1}^T$ and topic aspect-specific opinion model $\{\phi^{O,t}\}_{t=1}^T$. Parameter $y_{d,s,n}$ denotes the category of word, i.e. background word, feature word or sentiment word. Parameter $u_{d,s,n}$ indicates whether the word is general aspect or aspect-specific. The maximum model is used to train parameters $\pi^{d,s,n}$ and $x^{d,s,n}$. The word distribution $w^{d,s,n}$ is as follows:

$$\omega(d, s, n) \sim \begin{cases} \text{Multi}(\phi^B), & y_{d,s,n} = 0 \\ \text{Multi}(\phi^{A,Z,d,s}), & y_{d,s,n} = 1, u_{d,s,n} = 0 \\ \text{Multi}(\phi^{A,g}), & y_{d,s,n} = 1, u_{d,s,n} = 1 \\ \text{Multi}(\phi^{O,Z,d,s}), & y_{d,s,n} = 2, u_{d,s,n} = 0 \\ \text{Multi}(\phi^{O,g}), & y_{d,s,n} = 2, u_{d,s,n} = 1 \end{cases}$$

The author's sentiment orientation for comment target can be determined by previous

semantic rule-based techniques according to sentiment polarity of comment words. The major advantage of this model is that it can improve algorithm efficiency by extracting comment targets and comment words at the same time.

Besides, there are many topic model-based methods. For example, in 2011, Sauper Christian et al. proposed a joint topic-sentiment model to detect sentiment orientation for short documents ^[20], also known as HMM-LDA model as the Hidden Markov Model is used therein. In 2012, Mukherjee Arjun and Bing Liu proposed a hybrid model based on semi-supervised learning which allows users to provide seed words so as to improve the precision of sentiment analysis ^[21].

6.3 Social Network Sentiment Analysis Techniques

The new feature of online social network brings some new problems for traditional long text sentiment analysis technique, and further creates some sentiment analysis techniques for online social network. For example, the sentiment analysis technique specially for short text, the sentiment analysis technique using the mutual influence between groups in social network, a series of data processing technology for possible influence on true sentiment analysis by spam users and opinions, etc. This section introduces some research work of the above work.

6.3.1 The Sentiment Analysis Technique for Short Text

With the rapid development of Twitter, Facebook, Sina Weibo, people could post their views and opinions in the network anytime anywhere. Different from long text such as traditional news and reports, text in social network has short length, irregular grammar and lots of noise. Therefore, the research of sentiment analysis technique for short text in social network is of great importance.

In 2009, Go Alec et al. tested the sentiment classification effect of supervised learning algorithm on Twitter short text with such models as multinomial Bayesian classification, maximum entropy and support vector machines model. Go Alec et al. adopted emoticon in Twitter, rather than the method of obtaining training set by manual labeling in long text, to get positive and negative comment, thereby saving lots of manual labeling cost and significantly improving the size of training set. Through Twitter API, they collected the microblogs containing “:)” as positive comments and “:(” as negative comments, with

pre-processing including removing username, URL and repetitions. Single-factor model, two-factor model and mixed model were ultimate test methods in feature selection. Their sentiment classification result is about 80%, which is close to supervised learning method used by Pang Bo et al. in Reference [13].

Pak Alexander and Patrick Paroubek also adopted emoticon to get training set. They added objective information into subjective emoticon ^[23], thereby extending model to three-factor classification of traditional sentiment analysis. For the sentiment classification method, they adopted Naïve Bayes classification to obtain initial results and used information entropy to remove the influence of n -gram to improve classification to improve classification results.

Many social network media open their API (Application Programming Interface) for users to read content, such as Twitter, Facebook, Sina Weibo and Tencent Weibo. Users could capture data as required by search interface provided by social network media. Collecting data by emoticon saves lots of manual labeling work and greatly improves training set size. The following are examples for obtaining sentiment training set by API of Twitter and Sina Weibo.

Example 6-7 Collecting sentiment document training set by API.

1) Twitter API (<https://dev.twitter.com>)

Twitter API provides users with a series of search rules for the convenience of user's quick search. The positive document could be obtained by “:)” and the negative document could be obtained by “:(” . Users could also obtain the required data set by combining rules with text content, for example:

movie -scary :)	Containing “movie”, but not “scary”, and with a positive attitude.
flight :(Containing “flight” and with a negative attitude.

2) Sina Weibo API (<https://open.weibo.com>)

Sina Weibo provides lots of emoticons and specific emoticon API for users (<http://open.weibo.com/wiki/2/emotions>). In Sina Weibo, the emoticon posted by users is transformed to corresponding text. Regular expression “[**]” are used for labeling, such as [happy], [sad], [joy]. Users could adopt multiple key words to obtain microblog including specific emoticons.

Although the usage of emoticon saves the cost of manual labeling, it also brings noise and reduces the accuracy of training data. In 2012, Liu Kunlin et al. analyzed the influence of collecting training set by manual labeling and emoticon on sentiment analysis results ^[24]. Assume the training set collected by manual labeling as A , and the training set collected by

emoticon as B . They adopted probabilistic model for text modeling and calculated the sentiment classification probability of feature key words in A and B separately. Calculate the final sentiment classification results by Laplacian smoothing.

$$P_{co}(w_i | c) = aP_a(w_i | c) + (1 - a)P_u(w_i | c)$$

wherein $p_a(w_i | c)$ denotes the sentiment classification probability of feature word w_i in manual labeling training set, $p_u(w_i | c)$ denotes the sentiment classification probability of feature word w_i in the emoticon training set and a denotes smoothing factor. The experimental results show that the combination of two training sets could improve the sentiment classification accuracy.

With the emerging of microblog, the sentiment analysis for short text becomes important in social network sentiment analysis area. Nowadays, many conferences focus on sentiment analysis. For example, conferences such as NLP&CC and COAE regard sentiment analysis as an important part. Sentiment evaluation is different from the traditional three-factor sentiment classification (positive, neutral and negative), and adopts more fine-grained model. For example, NLP&CC2013 classifies users' sentiment into 7 categories (anger, disgust, fear, happiness, like, sadness and surprise). Zhang Lumin et al. adopt sentiment vector model to express the users' diversified sentiment in social network, and build hierarchical structure of sentiment vector based on cluster^[25]. Steps of algorithm are as follows:

Algorithm 6-4 Hierarchical sentiment vector model.

Step 1: Combine with sentiment check list in clinical psychology to exact initial sentiment vectors which could well express sentiment.

Step 2: Monitor microblog data stream and automatically discover and absorb new cyber words which could express sentiment based on statistics according to large-scale corpus. Establish self-learning and self-update mechanism of sentiment vector to guarantee the comprehensiveness.

Step 3: Use bottom-up method to establish hierarchical structure based on classification and summary. Label bottom sentiment vector and establish orientation analysis layer based on orientation word.

The ultimate hierarchical sentiment model is as shown Figure 6-3.

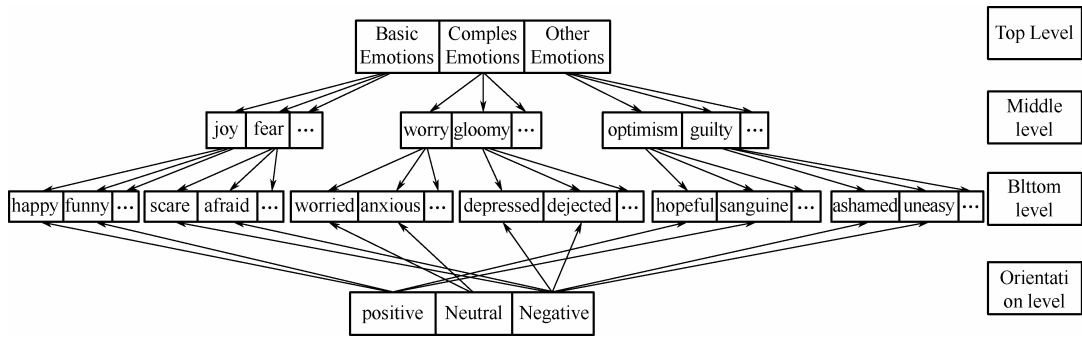


Figure 6-3 Hierarchical sentiment vector model

Microblog sentiment expression model based on sentiment vector could effectively express diversified sentiment. Use the method combining with clinical psychology to construct sentiment vector and the self-update mechanism to guarantee the comprehensiveness and authority. The hierarchical structure constructed by bottom-up method could avoid sparsity.

For sentiment analysis of topic, Wang Xiaolong et al. analyze sentiment by constructing hashtag-graph at topic layer for hashtag in Twitter^[26]. Assume $HG = \{H, E\}$ indicate hashtag-graph wherein $\forall h_i \in H$ denotes a hashtag and $e_k = \{h_i, h_j\} \in E$ denotes that h_i and h_j appear in the same tweet.

Example 6-8 Hashtag-graph model (See Figure 6-4).

In the hashtag-graph model, hashtag could be classified into three categories: topic tag, such as #president and #healthcare; sentiment tag such as #ideal and #leader; sentiment-topic tag, such as #iloveobama. The link between two hashtags indicates they appear in the same microblog.

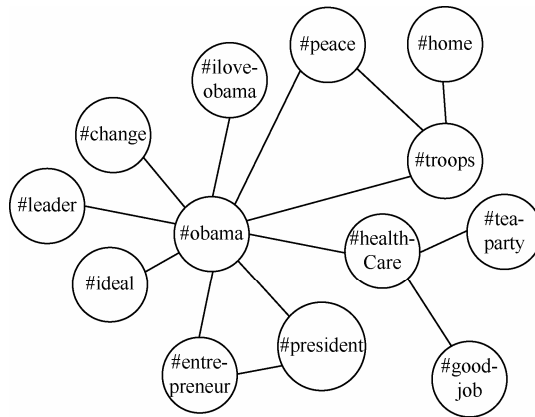


Figure 6-4 Hashtag-graph model

Given a hashtag-graph model HG , the main task is to label the sentiment orientation $y_i \in \{\text{pos}, \text{neg}\}$ for each hashtag $h_i \in HG$. According to Markov assumption, the sentiment polarity of a hashtag is determined by the polarity of microblog containing the hashtag as well as its neighbors. Therefore, this problem can be transferred to an optimization function:

$$\log(P(y | HG)) = \sum_{h_i \in H} \log(\phi_i(y_i | h_i)) + \sum_{(h_i, h_j) \in \mathcal{E}} \log(\psi_{i,j}(y_j, y_k | h_j, h_k)) - \log Z$$

wherein Z is the normalization factor, and function ϕ and ψ are defined as follows:

$$\begin{aligned} \phi_i(y_i | h_i) &= \sum_{\tau \in T_i} P_{y_i}(\tau) \\ \psi_{i,j}(y_j, y_k | h_j, h_k) &= \frac{\#(h_j, h_k)}{\#(h_j) + \#(h_k)} \cdot I_{y_j=y_k} \end{aligned}$$

wherein $\phi_i(y_i | h_i)$ denotes the probability that hashtag h_i is the sentiment orientation y_i ; $\#(h_j, h_k)$ denotes the number that h_j and h_k appear at the same time; $\#(h_j)$ denotes the number that hashtag h_j appears alone; $I_{y_j=y_k}$ is a decision-making function taking 1 if $y_j = y_k$ and 0 otherwise. Topic hashtag can be classified according to sentiment classification results of final hashtag.

In 2010, Bermingham and Smeaton compared the effectiveness of SVM algorithm and multinomial Bayesian algorithm between long text and short text^[27]. The results showed that SVM got better performance for long text data set while multinomial Bayesian algorithm got better performance for short documents. At the same time, the results showed that, although short text in social network like Twitter contains massive noise, it is easier to perform sentiment analysis for short text than long text.

6.3.2 Sentiment Analysis Based on Collective Intelligence

In social networks, users can express their views and opinions at will, but are unconsciously influenced by other nodes in social network based on the link structure. Interaction function provided by social network enhances sentiment interaction between users, and enable sentiment information to diffuse with the structure of social networks.

Thelwall Mike performed sentiment analysis on friend relationship network in Myspace, and found that linked users are more likely to have the same sentiment orientation^[28]. Bollen Joham et al. researched the homogenesis phenomenon of happiness in social network based on massive data from Twitter in 2011, and found that users are

more likely to select friends with the same happiness exponent ^[29]. They constructed the graph based on mutual follow relationship, and used Jaccard similarity to calculate the weight of edges as follows.

$$w_{ij} = \frac{\|C_i \cap C_j\|}{\|C_i \cup C_j\|}$$

Then, based on document dataset and corpus dataset, they quantified user's happiness exponent, and use Pearson index to calculate the correlation between vectors $S(S)$ and $S(T)$, which are constructed by grouping SWB value of starting point and end point of all edges.

Tan Chenghao et al. researched user level sentiment analysis techniques in Twitter^[30]. For each topic, they constructed a heterogenous graph model based on linked structure and content information as shown in Figure 6-5.

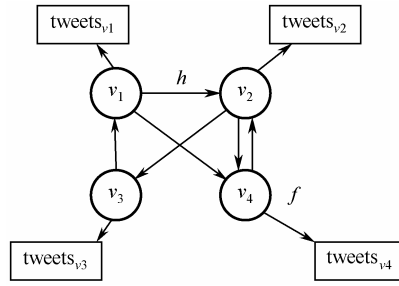


Figure 6-5 Heterogenous graph model based on linked structure and content information

wherein v_i denotes the user node, and tweets_{v_i} denotes the content of microblog posted by user. h denotes the link structure between users, and f is the link structure between user and their microblogs. For each user v_i , the sentiment label y_i denotes the sentiment orientation. Based on Markov assumption, the sentiment label of a user is determined by the sentiment in his/her microblogs and of his/her neighbors. Therefore, the model function is as follows:

$$\log P(Y) = \left(\sum_{v_i \in V} \left[\sum_{t \in \text{tweets}_{v_i, k, t}} \mu_{k, t} f_{k, t}(y_{v_i}, \hat{y}_t) + \sum_{v_j \in \text{Neighbors}_{v_i, k, t}} \lambda_{k, t} h_{k, t}(y_{v_i}, y_{v_j}) \right] - \log Z \right)$$

wherein $\mu_{k, t}$ and $\lambda_{k, t}$ is the weight factor. Function f denotes user-microblog factor, h denotes user-user factor with the following definitions:

$$f_{k,t}(y_{v_i}, \hat{y}_t) = \begin{cases} \frac{\omega_{\text{labeled}}}{|\text{tweets}_{v_i}|}, & y_{v_i} = k, \hat{y}_t = l, v_i \text{ labeled} \\ \frac{\omega_{\text{labeled}}}{|\text{tweets}_{v_i}|}, & y_{v_i} = k, \hat{y}_t = l, v_i \text{ unlabeled} \\ 0, & \text{otherwise} \end{cases}$$

$$h_{k,t}(y_{v_i}, y_{v_j}) = \begin{cases} \frac{\omega_{\text{relation}}}{|\text{Neighbors}_{v_i}|}, & y_{v_i} = k, y_{v_j} = l \\ 0, & \text{otherwise} \end{cases}$$

As not all users in social network contains labels, the authors adopted semi-supervised learning algorithm and used labeled dataset to perform label estimation based on the following four relationship networks.

- (1) Follow relationship network;
- (2) Bi-follow relationship network;
- (3) Mention relationship network;
- (4) Bi-mention relationship network.

The results showed that, compared with sentiment analysis only by text content, reasonable use of link structure information in social network can effectively improve the accuracy of sentiment analysis.

Besides, Zafarani Reza et al. systematically researched the sentiment influence between users in Live Journal ^[31]. Assume \wedge_s as the user set, $m(u, t)$ as the sentiment of user u at time t , $m(U, t)$ as the sentiment of a user group U at time t , the, the sentiment diffusion in social networks can be defined as given user group U at time t_i (U has the same sentiment), and user group U at t_j influences target user u at time t_j if

$$|m(U, t_i) - m(u, t_j)| \leq |m(\wedge_s, t_i) - m(u, t_j)| + b_1$$

and

$$|m(U, t_i) - m(u, t_j)| \leq |m(U, t_i) - m(u, t_i)| + b_2.$$

They used <excellent, poor> and <happy, sad> as the sentiment reference words, and used Google distance to determine the polarity for each feature word. They found that users posting relatively few microblogs are more likely to be influenced by others.

6.3.3 Mining Techniques on Spam Opinions in Social Network

In social network, spam users and internet marketers publish a lot of false information to push up product sales or make an event popular. Therefore, detection and analysis of spam opinions is significant for extracting factual information.

Jindal Nitin and Liu Bing proposed the concept of spam comment detection^[32] for products on the basis of sentiment analysis, which regarded spam comment detection as two-factor classification problem and used Logistic regression mode to classify user comments into spam comments and non-spam comments based on Amazon's 5.8 million product comments and their overlapping ratio. Furthermore, they took a more deep analysis on spam comments in 2008, and regarded content, authors, and objectives of comments as basic characteristics to effectively mine a lot of spam comments as false information from mass of comments.

For product rating behaviors (e.g., product quality divided into 1 ~ 5 levels), Lim Ee-Peng et al. found that spam users often comment on specific-concerned products and their comment results have a great deviation with the normal comments. Furthermore, they proposed spam users detection method based on comment objectives and spam users detection method^[33] based on deviation. For spam users detection method based on rating objectives, comment function on the basis of rating behaviors is as follows:

$$C_{p,e}(u_i) = \frac{s_i}{\text{Max}_{u_i' \in U^{s_i}}}$$

wherein u_i denotes user, p denotes products and s_i denotes non-standardized spam users comment function. The definition is as follows:

$$s_i = \sum_{e_{ij} \in E_{ij}, |E_{ij}| > 1} |E_{ij}| \cdot \text{sim}(E_{ij})$$

wherein E_{ij} denotes comment set of user i on product j and $\text{sim}(\)$ denotes similarity function defined on comment set.

Comment function for defining spam users on the basis of comments is as follows:

$$C_{p,v}(u_i) = \frac{s_i'}{\text{Max}_{u_i' \in U^{s_i'}}}$$

wherein

$$s'_i = \sum_{v_{i,j} \in V_{i,j}, |V_{i,j}| > 1} |V_{i,j}| \cdot \text{sim}(V_{i,j})$$

Based on the two comment functions mentioned above, the comment function for defining spam users is as follows:

$$C_p(u_i) = \frac{1}{2}(C_{p,e}(u_i) + C_{p,v}(u_i))$$

For spam users detection method based on deviation, spam users usually speak highly of concerned products and poorly on other products to achieve the purpose of promoting the sales volume. Therefore, the basis deviation of user comment is as follows:

$$d_{ij} = e_{ij} - \text{Avg}_{e \in E_{*j}} e$$

wherein e_{ij} denotes the comment of user i on product j , $\text{Avg}_{e \in E_{*j}} e$ denotes the average comment of all users on product j . Therefore, the final comment function for defining the average deviation of all comment behaviors of users is as follows:

$$c_d(u_i) = \text{Avg}_{e_{ij} \in E_{*j}} |d_{ij}|$$

Use the above two user behavior characteristics mentioned above to detect spam comments.

6.4 Extension and Transformation of Sentiment Analysis Technique

In social network, there are many other extensions and transformations related to sentiment analysis issue. This section will briefly introduce sentiment summary technique and sentiment analysis of interdisciplinary transfer learning. Please refer to references for more development of sentiment analysis.

6.4.1 Sentiment Summary Technique

Sentiment summary technique aims at automatically analyzing and concluding results of sentiment analysis for a large number of theme sentiment document, thus saving the time that user spend in reading relevant documents. There is a big difference between sentiment summary technique and traditional multi-document summary technique. The main purpose of traditional

multi-document summary is to extract the topics and their main contents from multiple documents, while the sentiment summary technique is based on sentiment comment object, aims at sentiment information conclusion for a certain topic or product and has obvious quantitative characteristics. For example, for a product, 80% documents are positive and 20% are negative.

In 2005, Liu Bing proposed the sentiment summary technique^[6] and displayed different product attributes in a structural manner. Users can compare features of different products horizontally so as to make a final decision based on demand. For a product set $P = \{P_1, P_2, \dots, P_n\}$, P_i denotes a kind of product. For each product P_i , $R_i = \{r_1, r_2, \dots, r_k\}$ denotes the corresponding comment set. For each feature f , if r_i contains feature f , then f is called as the explicit feature. If r_i does not contain feature f directly but implicitly, then f is called as the implicit characteristics. For each feature f , P_{set} denotes the positive comment set and N_{set} denotes the negative comment set. Therefore, sentiment summary tasks can be defined as follows: For each kind of product P_i and its comment set R_i , mine the explicit and implicit features from comments and extracting positive and negative comments of each feature f .

Example 6-9 Sentiment summary visualization.

This method of sentiment summary visualization rates different attributes of product and generates different sentiment summaries for different attributes. In Figure 6-6, phones provides several factors, such as sound, screen, battery, size, weight, etc. According to the user's comments, mine the sentiment orientation degree of different attributes and making a standard display.

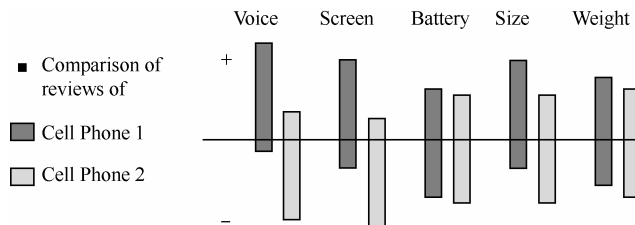


Figure 6-6 Sentiment summary structural display

In extraction technique of sentiment summary, Liu Bing extracted the product features based on supervised association rule mining. Firstly, label training set of feature manually; then mine frequent terms mined in data set based on the method of association rule mining and perform further manual processing to improve the accuracy of feature selection. In this process, the implicit features are mapped to the explicit features. Finally, combine

synonymous features based on Wordnet semantics to generate the final feature set. For each feature, display the proportions of positive and negative information according to statistical data.

Lu Yue et al. proposed a sentiment summary method^[34] based on online ontology for topics and different facets. They assumed that online ontology contained topics features and focus on solving two questions: how to choose effective features from lots of features and how to sort the features for helping users to read features. For selection of features, they proposed the method based on the size of set, the method based on opinion coverage and the method based on conditional entropy. For feature sorting, they sorted features according to the appearing order of topics features in article. Defining feature set to be sorted as $A' = \{A_i\}$ wherein each feature A_i , corresponds to a series of associating subjective sentences $S_i = \{S_{i1}, S_{i2}, \dots\}$. The coherent order of feature A_i, A_j was defined as:

$$\text{Co}(A_i, A_j) = \frac{\sum_{S_{i,k} \in S_i, S_{j,t} \in S_j} \text{Co}(S_{i,k}, S_{j,t})}{|S_i| |S_j|}$$

wherein $\text{Co}(S_{i,k}, S_{j,t})$ was set as 1 if $S_{i,k}$ appears in the article prior to $S_{j,t}$, otherwise as 0. $|S_i|$ denotes the number of words in a sentence. Therefore, for a feature set A' , optimal coherent degree can be defined as:

$$\hat{\pi}(A') = \arg \max_{\pi(A')} \sum_{A_i, A_j \in A', A_i < A_j} \text{Co}(A_i, A_j)$$

wherein $A_i < A_j$ denotes that feature A_i is prior to feature A_j . As it is a NP-hard problem, use greedy algorithm to obtain the local optimal sorting of features.

6.4.2 Sentiment Analysis Technology Based on the Mechanism of Transfer Learning

Text sentiment analysis algorithm has a strong correlation between the domains, the same word has different sentiment orientations in different domains. Therefore, it is necessary to research interdisciplinary sentiment classification by the mechanism of transfer learning.

Transfer learning divides the data source into the source domain and target domain. Source domain usually has a large number of labeled data set and the target domain usually does not have or only has a small amount of labeled samples. Transfer learning aims to

apply the feature representations or models learnt from source domain to target domain directly through feature association between source domain and target domain. Transfer learning does not requires that training data and test data follow the same distribution, so that information can be effectively shared and transferred between similar domains or tasks, which is different from traditional machine learning methods.

The features can be divided into two categories: domain dependent feature word and domain independent feature word. If the feature ranks high in both source domain and target domain (such as word frequency), the domains have nothing to do with the feature word, which is called as domain independent feature word. If the feature is strongly representative in the source domain but weakly representative in the target domain, the feature word is domain dependent feature word. In 2006, Yang Hui et al. realized interdisciplinary sentiment analysis task^[35] based on simple strategy of transfer learning based on feature selection in TREC task. It selected the words which rank high in both product and film comment and classified the sentiment successfully for product comment domain based on 2,041 positive comments and 2,217 negative comments in the movie comment.

Blitzer John et al. researched domain sentiment analysis techniques^[36] based on different product categories (book, DVD, electronic products and kitchen utensils) of Amazon in 2007. They not only improved the accuracy of sentiment analysis based on transfer learning methods but also researched the dependency between source domain and target domain, i.e. how to obtain the best transfer learning effect on a given target domain by selecting source domain.

For feature selection, Blitzer John et al. proposed structural correspondence learning^[37] in 2006. Firstly, they selected m central features from the source domain and target domain and each of them has strong representation in both source domain and target domain, i.e. domain independent feature, and then trained the mapping θ from initial feature space to shared feature space based on these central features. In the mapping feature space, bigger inner product of vectors indicates higher similarity. Blitzer et al. proposed SCL-MI algorithm to improve the above algorithm in 2007. For feature selection, they considered not only the high-frequency domain independent feature words but also mutual information of feature words and labels in the source domain so as to enhance the effectiveness of feature selection and the accuracy of sentiment analysis in the target domain. They used \mathcal{A} -distance based on SCL mapping to measure the applicability of source domain in target domain, and only considered the feature words that lead to different

classification results but ignore other differences between the source and target domain. \mathcal{A} -distance of two probability distributions was defined as

$$d_{\mathcal{A}}(D, D') = 2 \sup_{A \in \mathcal{A}} |Pr_D[A] - Pr_{D'}[A]|$$

wherein sup denotes the upper bounding function.

6.5 Summary

With the rapid development of Internet technology, online social network is becoming main medium for users to express their views and diffuse information. In this chapter, we provide an overview of sentiment analysis and opinion mining in online social network, and systematically introduce sentiment analysis technique for long text like news report as well as the influence of link structure and group interaction features of social networking on sentiment analysis. Sentiment analysis for social network has great value for application and will play more and more important roles in financial, political, economic and other fields.

Currently, sentiment analysis technique for text is still under development and unified and mature theoretical system has not formed. Moreover, sentiment analysis technique for social network is still at the exploratory period, especially for sentiment analysis method for short text like Twitter. Online social network greatly enriches the corpus of sentiment analysis but brings more issues and challenges. We think the following issues in sentiment analysis technique for social network need to be further researched and explored.

(1) Sentiment model. Traditional text sentiment analysis uses three-factor model (positive, negative and neutral) to describe users' sentiment. However, in a social network, user sentiment is often diversified and much more complex especially on emergencies in social network. The traditional three-factor model is not sufficient for representing user sentiment. We need to establish sentiment model with higher granular degree in sentiment analysis for social network to represent the user sentiment more accurately. Although there are related researches (e.g., Reference [25]) and conferences (e.g., NLP&CC and COAT), it is necessary to further establish general and authoritative models.

(2) Sentiment analysis and group interaction. In social network, user sentiment is influenced by both its subjective consciousness and neighbor nodes. The emergence of online social network allows users to perform more frequent interaction and more common affective exchange. As existing work mainly focuses on analyzing the text sentiment orientation, it is necessary to further perform sentiment analysis work combined with social network features,

such as researching diffusion and diffusion model of sentiment in social networks, mining evolution law of opposed sentiment in groups, mining sentiment communities and so on.

References

- [1] Janyce M. Wiebe. Tracking point of view in narrative[C]. Computational Linguistics, 1994, 20(20): 223-287.
- [2] Sanjiv Das and Mike Chen. Yahoo! for Amazon: Extracting market sentiment from stock message boards[C]. In Proceedings of the Asia Pacific Finance Association Annual Conference(APFA),2001.
- [3] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews[C]. In Proceedings of WWW,2003: 519-528.
- [4] Pang Bo, and Lillian Lee. Opinion mining and sentiment analysis[M]. Foundations and trends in information retrieval 2.1-2(2008): 1-135.
- [5] Liu Bing. Sentiment analysis and opinion mining[M]. Synthesis Lectures on Human Language Technologies 5.1(2012): 1-167.
- [6] Liu Bing, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web[C]. In Proceedings of the 14th international conference on World Wide Web, ACM, 2005: 342-351.
- [7] Wilson Theresa, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis[C]. In Proceedings of hlt/emnlp on interactive demonstrations, Association for Computational Linguistics, 2005: 34-35.
- [8] Devitt Ann, and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach[C]. ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. 2007, 45(1): 984.
- [9] Kim Soo-Min, and Eduard H. Hovy. Crystal: Analyzing Predictive Opinions on the Web[C]. In EMNLP-CoNLL, 2007: 1056-1064.
- [10] Turney Peter D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.
- [11] Ding Xiaowen, Bing Liu, and Philip S. Yu. Aholistic lexicon-based approach to opinion mining[C]. In Proceedings of the conference on Web Search and Web Data mining (WSDM-2008).
- [12] Kim Soo-Min, and Eduard Hovy. Automatic detection of opinion bearing words and sentences[C]. In Companion Volume to the Proceedings of the International Joint Conference on Natural

Language Processing (IJCNLP), 2005: 61-66.

- [13] Pang Bo, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? : sentiment classification using machine learning techniques[C]. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
- [14] E. T. Jaynes. Information Theory and Statistical Mechanics. Phys. Rev. 106, 620 – Published 15 May 1957
- [15] Hang Li. Statistical learning method [M]. Beijing: TsingHua Press. 2012.
- [16] Hofmann, Thomas. Probabilistic latent semantic indexing[C]. In Proceedings of conference on uncertainty in artificial intelligence. 1999.
- [17] Blei David M. , Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation[J]. The Journal of machine Learning research 3(2003): 993-1022.
- [18] Bishop C.M. Pattern recognition and machine learning[M]. Vol.4.2006: Springer, New York.doi: 10.1117/1.2819119.
- [19] Zhao Wayne Xin, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid[C]. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010: 56-65.
- [20] Sauper Christina, Aria Haghighi, and Regina Barzilay. content models with attitude[C]. In Proceedings of the 49th annual meeting of the association for computational Linguistics. 2011.
- [21] Mukherjee Arjun and Bing Liu. aspect extraction through Semi-Supervised modeling[C]. In Proceedings of 50th annual meeting of association for computational Linguistics. 2012.
- [22] GoAlec, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision [C]. CS224N Project Report, Stanford(2009): 1-12.
- [23] Pak Alexander, and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]. LREC.2010.
- [24] Liu Kunlin, Wu Junli, and Minyi Guo. Emoticon Smoothed Language Models for Twitter Sentiment Analysis[C]. AAAI. 2012.
- [25] Zhang Lumin, Yan Jia, Bin Zhou, and Yi Han. Microblogging sentiment analysis using emotional vector[C]. IEEE International Conference on Cloud and Green Computing, 2012, 430-433.
- [26] Wang Xiaolong, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach[C]. In Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011: 1031-1040.
- [27] Bermingham Adam, and Alan F. Smeaton. Classifying sentiment in microblogs: is brevity an advantage? [C]. Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

- [28] Thelwall Mike. Emotion homophily in social network site messages[J]. First Monday 15.4(2010).
- [29] BollenJohan, Bruno Gonçalves, Guangchen Ruan, and Huina Mao. Happiness is assortative in online social networks[J]. Artificial life 17, no. 3 (2011): 237-251.
- [30] Tan Chenhao, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks[C]. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011: 1397-1405.
- [31] ZafaraniReza, William D. Cole, and Huan Liu. Sentiment propagation in social networks: a case study in livejournal[J]. Advances in Social Computing. Springer Berlin Heidelberg, 2010: 413-420.
- [32] JindalNitin and Bing Liu. mining comparative sentences and relations[C]. In Proceedings of national conf. on artificial intelligence. 2006.
- [33] Lim Ee-Peng, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady W. Lauw. Detecting Product review Spam users using rating Behaviors[C]. In Proceedings of acm International conference on information and Knowledge management (CIKM-2010), 2010.
- [34] Lu Yue, Huizhong Duan, Hongning Wang, and ChengXiang Zhai. Exploiting structured ontology to organize scattered online opinions[C]. In Proceedings of international conference on computational Linguistics, 2010.
- [35] Yang Hui, Luo Si, and Jamie Callan. Knowledge transfer and opinion detection in the trec 2006 blog track[C]. In Proceedings of trec, 2006.
- [36] Blitzer John, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification[C]. In Proceedings of annual meeting of the association for computational Linguistics, 2007.
- [37] Blitzer John, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning[C]. Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006.
- [38] He Yulan, Chenghua Lin, and Harith Alani. automatically extracting polarity-bearing topics for crossdomain sentiment classification[C]. In Proceedings of the 49th annual meeting of the association for computational Linguistics, 2011.
- [39] Wang, Guan, Sihong Xie, Bing Liu, and Philip S. Yu. identify online Store review Spam users via Social review graph[J]. acm transactions on intelligent Systems and technology, 2011.
- [40] Zhao Yanyan, Qing Bing, Liu Ting. Text sentiment analysis [J]. Journal of Software 2008, 21(8) : 1834-1848.
- [41] Zhou Lizhu, He Yukai, Wang Jianyong. Overview on sentiment analysis research [J]. Computer Application. 2008, 28(11) : 2725-2728.

Influence Analysis and Its Technologies

The influence of individuals in social networks had been widely studied by researchers from various fields, such as, social psychology, telecommunications, marketing, and computer science, which plays an important role in the guidance of public opinions and social operations. With the emergence of a large number of online social network services and the participation of users, researches on the influence of individuals in social networks have attracted the attention of many scholars at home and abroad. In the era of online social networking, social networks have exhibited a significant impact on people's daily life and behavior patterns. A small number of malicious users and opinion leaders are found to fabricate and disseminate public opinions by taking advantage of the social networking services. By expressing their opinions on the current events, and interacting with the media and the Internet users, opinion leaders are often very powerful in influencing the minds of their fans and the direction of public opinions. In recent years, opinion leaders has played an important role in participation and guidance of public opinions, in such events as "Cracking Down on the Abduction of Women and Children through Weibo", and offering "free lunch" for children in poor areas, while in other incidents like demolition, petition, accidents and disasters, opinion leaders have also played an important role in the generation, fermentation, propagation, and sensationalization of such events. At the same time, the analysis of individual influence has been widely used in a number of areas, such as recommendation systems, social network information dissemination, link prediction, viral marketing, public health, expert discovery, incident detection and advertising etc. Therefore, the analysis of individual influence in social networks has great theoretical value and practical significance.

How to identify users of high influence in a heterogeneous, multi-attribute social network, and analyze the strength of influence between users in social networks is one of the key issues in information related decisions making in the fast-changing age of the Internet.

The contents of this chapter are as arranged as follows: Section 7.2 introduces the methods for analysing and calculating the strength of influence between users, including the network structure-based, the behavior-based, and the topic-based strength calculation methods; section 7.3 describes the methods for identifying influencers, which involves the network structure-based individual influence calculation, the Pagerank algorithm, the behavior-based individual influence calculation, and the topic-based individual influence calculation.

7.1 Introduction

The technologies for analysing individual influence in social networks are mainly classified into two types: one is based on analysing the strength of influence between users, and the other is based on identifying the influencers.

1. Influence Strength Between Users

The strength of influence between individuals in a social network depends on many factors, such as, the network distance between users and the temporal action pattern. To put it in a simpler way, the strength of influence is the quantitative size of the edge of a social network. For example, in Figure 7-1, the strength of influence between v1 and v2 is 0.33, the strength of influence between v2 and v4 is 0.5.

2. Identification of Influencers

In order to identify influencers in a social network, the individual influence ranking technology is mainly applied. The influence of an individual depends on many factors, such as, the network structure, and the behavior pattern of users in the social network. Simply speaking, the influence of an individual is the quantitative size of the node in a social network. For example, in Figure 7-1, the influence of v1 is quantified as 0.025, and the influence of v2 is 0.0259.

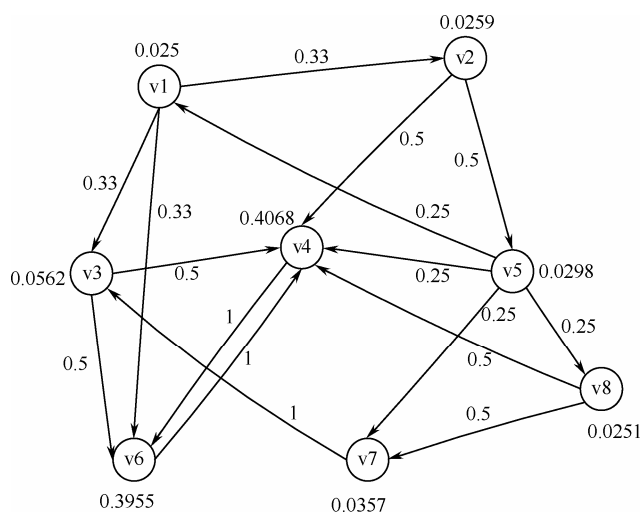


Figure 7-1 Analysis of individual influence in social networks

The concepts involving individual influence analysis technology employed in social network studies first appeared in sociology, which were then further explored in information science.

At first, researchers in the field of sociology explored the inequality of individual influences, and discovered the diversity of individual influences. In 1955, Professor Elihu Katz (1926—)^[1] proposed the two-level communication theory based on a study of voters' intention to vote during the US presidential election, in which he discovered the difference of individual influences, and found that a small part of the opinion leaders or influencers held the influencing power over the majority of ordinary people.

Later in 1962, Everett Rogers (1931—2004)^[2] defined the term “influential” or “influencer”: an individual who is able to change the idea of other individuals in a certain degree. Influencers usually have four characteristics: ①inclined to conveying their own ideas to others; ②representing the views of most common people; ③holding novel ideas; ④also known as opinion leaders, innovators, hubs, connectors, mavens, etc.

Early works had been focused on exploring and analyzing the performance and related factors of influence in social activities, in which the function model and the generation mechanism of social influence were studied profoundly. Many social phenomena associated with influence and the underlying principles were discovered. However due to the limited size of samples at that time, the data attainable was limited, which requires sufficient objective data support and verification.

With the rapid development of the Web2.0 technology and the rise of online social

networks, for the first time researchers have the opportunity to analyze and study the complex relationships among the mass of interactive information and large-scale social networks. The influence research turns to the rich data generated by online social networks for support, to establish various influence analysis models and make extensive use of different quantitative techniques; researchers have also carried out research and exploration on the influences of users themselves, the mutual-influence of users during online interactions, the interaction between users and their societies, and the evolution of influence over time. These researches have not only validated and expanded many of the early assumptions and theoretical models, but also observed many interesting phenomena and rules. In addition, a large number of scientific research issues and application scenarios associated with social influence have been discovered in the new research environment.

In 2006, Noah Friedki^[3] defined “social influence”, and pointed out that the existing social networking will change the (variable) characteristics of people.

Finally, in order to measure the ability of social influence quantitatively, we defined influence strength as the quantitative social influence, which indicates the degree of interaction between individuals on social networks, also known as Strength of Relationship, Relationship Strength, or Tie Strength.

Definition 7-1 (influence strength): Given two user nodes u and v in the network $G = (N, E)$, we denote $I_u(v) \in R$ as the influence strength of user v on user u . Furthermore, if $e_{uv} = 1$, we call $I_u(v)$ the direct influence of user v on u ; if $e_{uv} = 0$, we call $I_u(v)$ the indirect influence of user v on u . The influence strength satisfies the anisotropy, i.e., $I_v(u) \neq I_u(v)$.

Overall, online social influence analysis mainly involves two aspects: ① Influence strength measurement. On the basis of the qualitative identification of social influence, in the face of complex social relations, how to design and choose a measurement method that not only has certain universality, but also can fully explore the characteristics of social networks is one of the core issues in the field. ② Influentials identification. On the basis of quantitative calculation of social influence, how to accurately calculate the individual's influence and identify influentials based on ranking is of significant value for the analysis of social network evolution, social behavior, information transmission mode and many other issues.

7.2 Influence Strength Calculation

Early researches on the influence strength of individuals in social networks only

considered the network structure, and simply used the overlap of common neighbors, the edge betweenness, the frequency of forwarding and other factors to measure influence strength. Later, in order to address the problem of limited use of the information on user behaviors and interactions based on the method of network topology, some researchers put forward a method for calculating influence strength based on individual behaviors. In view of the difference in users' influence strength in different topics, some researchers studied the topic-based influence strength calculation.

7.2.1 Influence Strength Calculation Based on Network Structure

1. Influence strength calculation based on the overlap of neighborhoods

In 1973, Mark S. Granovetter^[4] proposed the method of measuring the influence strength between two nodes based on the overlap of neighborhoods. Generally, if the overlap of neighborhoods between A and B is large, we consider A and B to have a strong influence strength. Otherwise, the influence strength is low. We formally define the strength $S(A, B)$ in terms of the Jaccard coefficient:

$$S(A, B) = \frac{|n_A \cap n_B|}{|n_A \cup n_B|} \quad (7-1)$$

Here, n_A and n_B denotes the sets of neighbors of A and B , respectively. If the overlap of neighborhoods between A and B is large, we consider A and B to have a strong tie. Otherwise, they are considered to have a weak tie. Similar to Jaccard coefficient, the social influence in the network is measured by overlapping similarity and cosine distance.

2. Influence strength calculation based on edge betweenness

In 1977, Linton C. Freeman^[5] proposed the concept of betweenness to measure the importance of an edge in a network, which is supposed that the information flows between node s and t are evenly distributed on the shortest paths. That is, if there is only one shortest path between them, such a path is given a weight of unity. The betweenness of an edge e is calculated by summarizing the total “weights” of all shortest paths going through it. We defined the influence strength as follows.

$$E^{\text{BET}}(e_{ij}) = \sum_{s < t} |g_{st}^{ij}| \quad (7-2)$$

Here, $|g_{st}^{ij}|$ denotes the number of the shortest paths between s and t simultaneously passing i and j (through the edge e_{ij}). The betweenness of an edge is calculated by summarizing the total number of all shortest paths going through the edge e_{ij} .

3. Influence strength calculation based on forwarding frequency

In 2006, when the propagation of the influence of blogs was analyzed in references, Directed Multigraph was used to represent the influence between nodes. In their work, directed graph with edge weights indicates how much influence a particular source node had on its destination, the direction of the arc represents the source of influence, the weight represents the intensity of the influence, The formula is given as where $c_{u,v}$ indicates number of arcs from u to v and $\deg^{\text{in}}(v)$ denotes the in-degree of v :

$$w_{u,v} = \frac{c_{u,v}}{\deg^{\text{in}}(v)} \quad (7-3)$$

How to use the above formula to calculate the influence of the individual in blog post networks was elaborated in Figure 7-2.

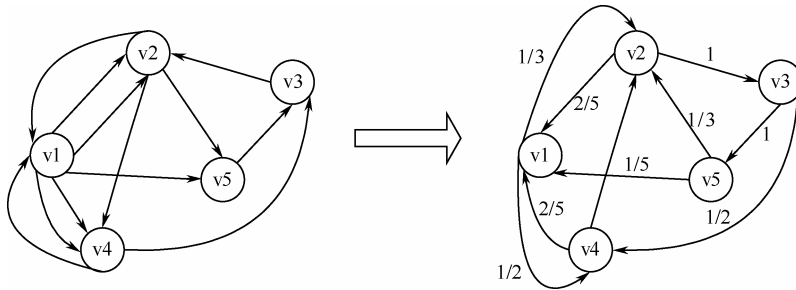


Figure 7-2 Influence strength calculation based on forwarding frequency

7.2.2 Behaviour-based Influence Strength Calculation

However, influence calculation based on network topology has some inherent defects: first, the social network topology that researchers achieved is static, equivalent to a snapshot of the original network, on which all explicit social relations before the acquisition are recorded; that is to say, all the connections established ten years ago and those at one second ago are collected at the same time; the connection that received just one notice and the connection of intimate communications between two friends are treated equally in the calculation model; secondly, in such a network topology, the

weights of all connections are equal or identically distributed, meaning that the connected users have the same influence on each other, or that the influence among the users in the social network satisfies a simple probability function. The above situation is obviously inconsistent with the reality, and the fundamental reason is that the limited use of information on user's behaviors and their interactions in the calculation method based on network topology, which leads to the deviation of the results of this method from the actual situation.

User behaviors in online social network include information release; shopping, commenting on topics, forwarding information, establishing friendship, etc. Analysis of the distribution rules and causality of such behaviors will help to not only evaluate the influence between the initiator of the behavior and the disseminator, but also predict the behavior of people in social networks, and deepen our understanding of human social behavior.

In general, online social networks will record a large amount of information arising from people's interactive activities, which include all kinds of user behavior data. By analyzing these data, we can not only measure the influence between users and the way and the scope of its diffusion, but also establish the social relations between users. Goya Amit et al.^[7] studied the influence of both users and their behaviors based on logs:

$$\text{infl}(u) = \frac{|\{a \mid \exists v, \Delta t : \text{prop}(a, v, u, \Delta t) \wedge 0 \leq \Delta t \leq \tau_{v,u}\}|}{A_u} \quad (7-4)$$

$$\text{infl}(a) = \frac{|\{u \mid \exists v, \Delta t : \text{prop}(a, v, u, \Delta t) \wedge 0 \leq \Delta t \leq \tau_{v,u}\}|}{U(a)} \quad (7-5)$$

Where, u and v denote users respectively; a denotes action; Δt represents the time interval of actions; $\tau_{v,u}$ is a time constant; $\text{prop}(a, v, u, \Delta t)$ denotes the transmission of actions between users; A_u represents the number of actions of user u ; $U(a)$ denotes the number of users of executing actions. Different from the network topology based influence strength calculation method, the above model uses the frequency of action transmission as the indicator for measuring influence strength, and the execution scope of the action is used to measuring the influence of the action itself. In order to calculate the influence of the user on their neighbors, they designed the static probability model, continuous time model and discrete time model. The relative parameters were estimated by machine learning method and the influence coefficient between users in Flickr was

calculated.

Figure 7-3 shows the schematic diagram of the method for calculating influence strength based on behavior.

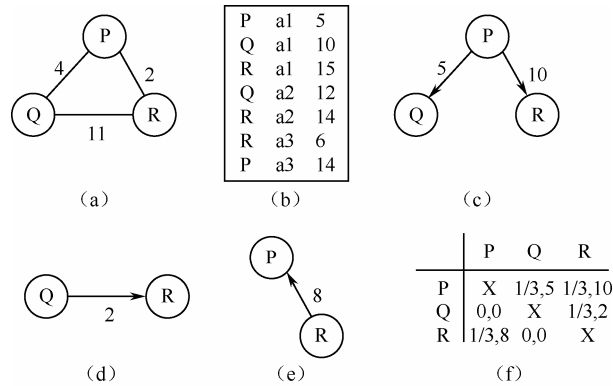


Figure 7-3 Influence strength based on the behavior

7.2.3 Topic-based Influence Strength Calculation

In social activities, information is often generated and transmitted in the form of topics. Studies found that users' influence strength varies with different topics. Therefore, measuring influence strength by using topics as the basic object is conducive to depicting influence between users from many different perspectives. In establishing an influence strength model, it is possible to construct the relationship between a user and a topic by directly resorting to the content of the topic and the degree of user's participation in the topic, without having to turn to the social network topology established based on such user behaviors as applying for a friend relation or following someone as the model input. The influence strength resulting from analysis based on the former method is named "implicit influence", while that of the latter is called "explicit influence". At the same time, the entities in online social networks include users, text, and multimedia information, which internally and mutually form a heterogeneous network structure which is more complex than homogeneous networks.

Tang et al.^[8] proposed a Topical Factor Graph (TFG) model to formulate the topic-level social influence analysis into a unified graphical model. In particular, the

motivation of the model is to simultaneously capture information related to topics, such as topical distributions, similarity between users' topics, and network structure. Factor graph model was based on object likelihood function:

$$P(\mathbf{v}, \mathbf{Y}) = \frac{1}{Z} \prod_{k=1}^N \prod_{z=1}^T h(y_1, \dots, y_N, k, z) \prod_{i=1}^N \prod_{z=1}^T g(v_i, y_i, z) \prod_{e_{kl} \in E} \prod_{z=1}^T f(y_k, y_l, z) \quad (7-6)$$

Where, $\mathbf{v} = [v_1, \dots, v_N]$ denotes a set of observed variables; $\mathbf{Y} = [y_1, \dots, y_N]$ denotes a set of latent variables; g and f respectively represent feature functions of nodes and edges; h denotes global feature functions; and Z represents normalized factors.

Figure 7-4 shows the schematic diagram of the factor graph model, in which $g(\cdot)$ represents a feature function defined on a node; $f(\cdot)$ represents a feature function defined on an edge; $h(\cdot)$ represents a global feature function defined for each node; they indicate forwarding frequency and other factors; the latent vector y_i models the topic-level influence from other nodes to node v_i .

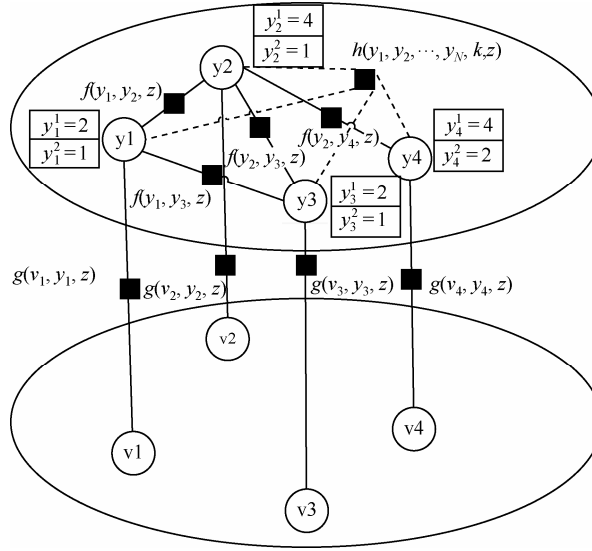


Figure 7-4 Schematic diagram of the factor graph model

Liu et al.^[9] proposed a generative graphical model which leveraged both heterogeneous link information and textual content associated with each user in the network to mine the implicit influence between users and predict user behaviors by making use of the similarity of textual content. The generative probabilistic model is as follows(see

Figure 7-5).

By using Gibbs sampling and iterative learning, the influence strength $\gamma_d(c)$ on d from c due to its citation of c is:

$$\gamma_d(c) = \frac{1}{K} \sum_{k=1}^K \frac{C_{d,c,s}(d, c, 0)^{(k)} + \alpha_\gamma}{C_{d,s}(d, 0)^{(k)} + |L(d)| \cdot \alpha_\gamma} \quad (7-7)$$

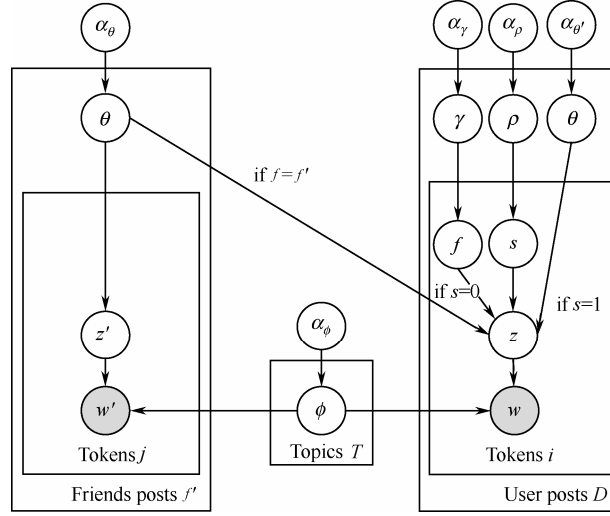


Figure 7-5 Generative model

For random variables $\text{var}_1, \text{var}_2, \dots, \text{var}_n$, the notation $C_{\text{var}_1, \text{var}_2, \dots, \text{var}_n}(\text{val}_1, \text{val}_2, \dots, \text{val}_n) = |\{ \forall i : \text{var}_{1,i} = \text{val}_1 \wedge \text{var}_{2,i} = \text{val}_2 \wedge \dots \wedge \text{var}_{n,i} = \text{val}_n \}|$ counts occurrences of a configuration $\text{val}_1, \text{val}_2, \dots, \text{val}_n$. For example, $C_{d,c,s}(1, 2, 0)$ denotes the number of tokens in document 1 that are assigned to citation 2, where the coin result of Bernoulli distribution is 0. α_γ represents the superparameter (Bernoulli distribution parameter); K represents the number of iterations; $|L(d)|$ indicates the length of the document. As shown in Figure 7-6, the “Bryant” vocabulary of user A comes from its friends B and C, and according to the probability-generating model, it is possible to calculate the probability of “Bryant” of user A respectively from its friends B and C.

This method integrates the information of user interaction in the social network structure, and by analyzing the relationship between topic information and users, it can measure both the generation and the change process of user influence more accurately.

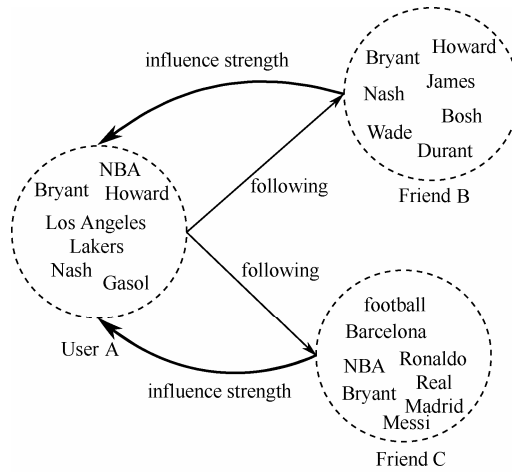


Figure 7-6 Influence strength based on generative model

7.3 Identification of Influentials

The identification of influentials is mainly realized by using the technology of ranking individual's influence. The influence of individuals depends on many factors such as the structure of networks, and the behavior patterns of users in social networks.

The early methods of individual influence calculation are mainly based on such concepts as “degree centrality”, “closeness”, and “betweenness”. Later, some methods based on random walk, such as, HITS^[10] and PageRank^[11] were proposed to discover influentials. In recent years, with the increasing complexity of individual behavior and the diversity of topics in social networks, researchers have studied individual identification methods based on individual behaviours and topic levels.

7.3.1 Individual Influence Calculation Based Network Structure

Degree centrality: degree centrality refers to the number of nodes in a social network; that is, the number of nodes directly connected to a specific node, which measures the average influence of a node to its neighbors. Let A be the adjacency matrix of a network, and $\deg(i)$ be the degree of node i . The degree centrality of node i is c_i^{DEG} , i. e., the degree of the node is:

$$c_i^{\text{DEG}} = \deg(i) \quad (7-8)$$

For example, in Figure 7-1, the in-degree of node v_1 is 1, and the in-degree of node v_4 is 5.

Cha Meeyoung et al. ^[12] calculated the degree centralities of the following, forwarding and referring networks respectively, to measure the individual influence in Twitter. They used Spearman's rank correlation coefficient as a measure of the strength of the association between two rank sets.

$$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{N^3 - N} \quad (7-9)$$

Here, x_i and y_i are the ranks of users based on two different influence measures in a dataset of N users. The Spearman rank correlation coefficient is used to determine the degree of closeness between the two sets of rankings. The calculation is based on the rankings. The higher the consistency of the two groups, the higher the Spielman rank correlation coefficient. When the group variable is exactly the same, the Spearman rank correlation coefficient is 1.

Closeness: closeness is the sum of the short distances (shortest paths) between an individual and all other nodes in a social network. Closeness centrality can be used to measure a node's indirect influence to other nodes, or the distance from a node to others. It can also be used to measure the strength of a user's social ties. The higher the closeness centrality of a user, the shorter the distance between it and other users, and the faster its influence will be spread to other users. The most popular centrality measure in this group is the Freeman's closeness centrality. The closeness centrality c_i^{CLO} of node i is defined as follows.

$$c_i^{\text{CLO}} = e_i^T S \mathbf{1} \quad (7-10)$$

Here, S is a matrix whose (i, j) th element contains the length of the shortest path from node i to j , and $\mathbf{1}$ is the all-one vector.

The averages of the shortest distances to all other nodes are computed, and the computational load is relatively heavy. The advantage is that it can measure the indirect influence of a node.

Betweenness: betweenness refers to the ability of a node to be on the shortest path to

other nodes in a social network. It is used to analyze the influence of a node on the diffusion of information; that is, the extent to which the individual is between others, and whether it plays the role as an “intermediary”. The betweenness centrality c_i^{BET} of node i is defined as follows:

$$c_i^{\text{BET}} = \sum_{j,k} \frac{b_{jik}}{b_{jk}} \quad (7-11)$$

Here, b_{jk} is the number of shortest paths from node j to k , and b_{jik} is the number of shortest paths from node j to k that pass through node i . The naive algorithm for computing the betweenness involves all-pair shortest paths. This requires $O(n^3)$ time overhead and $O(n^2)$ space overhead. Brandes Ulrik^[13] designed a faster algorithm with the use of single-source-shortest-path algorithms. This requires $O(n+m)$ space overhead and runs in $O(nm)$ and $O(nm + n^2 \log n)$ time overhead, where n is the number of nodes and m is the number of edges.

The disadvantage of betweenness centrality is that the computational load is relatively heavy. The advantage of it is that this method can span structural holes in a social network.

HITS: This method was proposed by Jon Kleinberg^[10], also called as Hypertext Induced Topic Search. First, HITS was first used in search engines, where the importance of pages was measured by Hub and Authority. For given a node v_i in a network, the authority is defined as $a(v_i)$, and the hub is defined as $h(v_i)$. Hub and Authority are measured as follows.

$$a^{(k+1)}(v_i) = \sum_{v_j \in \text{inlink}[v_i]} h^{(k)}(v_j), \quad h^{(k+1)}(v_i) = \sum_{v_j \in \text{outlink}[v_i]} a^{(k+1)}(v_j) \quad (7-12)$$

Authority page: when a web page is often cited, it may be very important; even though a web page is not often cited, but if it is cited by important web pages, it is possible that it is also important; the importance of a web page is evenly passed to the citing web pages. Such a web page is called an “authoritive” page.

Hub page: A web page that provides a link to an authoritative web page, which itself may not be important, or in other words, there is not many pages linking out to it, but it provides a collection of links to a site which is critical for a topic; for example, the list of references on the home page of a course.

In HITS Algorithm, both the authority and the hub are measured for each page. The HITS Algorithm is as follows.

```
initialize authority and hub weights, a0 and h0
```

```
while (not converged)
```

```
for each vertex i
```

$$a^{(k+1)}(v_i) = \sum_{v_j \in \text{inlink}[v_i]} h^{(k)}(v_j)$$

$$h^{(k+1)}(v_i) = \sum_{v_j \in \text{outlink}[v_i]} a^{(k+1)}(v_j)$$

```
end
```

```
end
```

Figure 7-7 is a schematic diagram of HITS. The authority of node v_1 is measured by the hub of v_2, v_3, v_4 . The hub of v_1 is measured by the authority of v_5, v_6, v_7, v_8 .

$$a(v_1) = h(v_2) + h(v_3) + h(v_4)$$

$$h(v_1) = a(v_5) + a(v_6) + a(v_7) + a(v_8)$$

Romero Daniel M et al ^[14] devised a general model for analysing influence in Twitter using the concept of passivity in a social network and developed an efficient algorithm similar to the HITS algorithm, called Influence-Passivity (IP) algorithm. For every

arc $e = (i, j) \in E$, they defined the acceptance rate $u_{ij} = \frac{w_{ij}}{\sum_{k:(k,j) \in E} w_{kj}}$. The acceptance rate can

be viewed as the dedication or loyalty user j has to user i . On the other hand, for every

$e = (j, i) \in E$, they defined the rejection rate by $v_{ji} = \frac{1 - w_{ji}}{\sum_{k:(j,k) \in E} (1 - w_{jk})}$. The algorithm is

based on the following operations:

$$\begin{aligned} I_i &\leftarrow \sum_{j:(i,j)} u_{ij} P_j \\ P_i &\leftarrow \sum_{j:(j,i)} v_{ji} I_j \end{aligned} \tag{7-13}$$

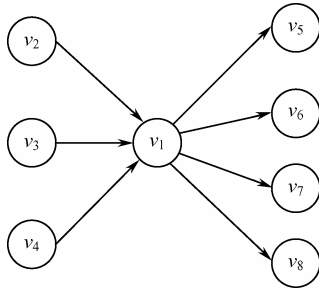


Figure 7-7 HITS

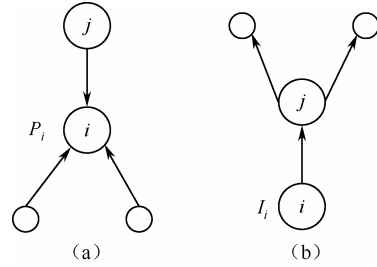


Figure 7-8 IP

For example, Figure 7-8 illustrates the IP algorithm. A user's influence score depends on:

- (1) the passivity score of its influence.
- (2) friends' rate of acceptance to its influence in relation to other users.

A user's passivity score depends on:

- (1) the influence score of its neighbors.
- (2) friends' rate of rejection against its influence in relation to other users.

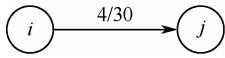


Figure 7-9 Weight calculation

The acceptance rate and the rejection rate are computed by using the above methods. The calculation of weights is critical. The weight of edge (i, j) is defined as the ratio of user i forwarding the posts of user j . Figure 7-9 gives an example of

weight calculation.

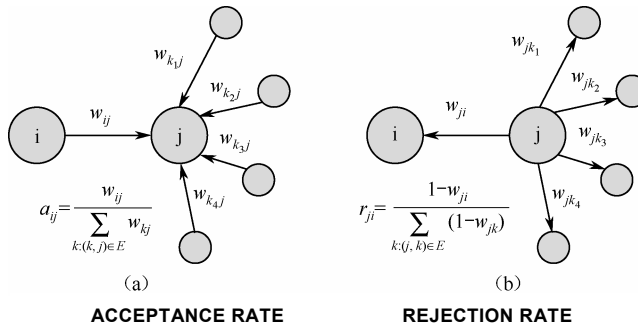


Figure 7-10 Diagram of acceptance rate and rejection rate calculations

7.3.2 PageRank

The authority and the hub were considered in HITS algorithm, whose values were computed by the iterative process. However, the repartition of influence was neglected.

PageRank was proposed by Larry Page^[11], which was taken in search engine. The influence of a page in PageRank was measured by values of linked-in pages, where the network structure was considered. The rating of a page is determined by the importance of all the links to it. Later, PageRank was applied to social networks by some researchers, where the influence of each node in social networks was measured. The Markov Model based random walk concept was adopted to simulate the behavior of browsing webpages. Let π be the score of influence. P be the transfer matrix of a social network, then PageRank is defined as follows:

$$\pi = \alpha P^T \pi + (1 - \alpha) \frac{1}{n} e, e = (1, 1, \dots, 1)^T \quad (7-14)$$

Here, α is a jump factor, $\frac{1}{n}e$ is a restart vector, assuming that each node is visited randomly by other nodes equiprobably. Let $i \rightarrow j$ be a direction from node i to node j . The adjacency matrix M of the social network G is defined as follows.

$$M(i, j) = \begin{cases} 1, & i \rightarrow j \\ 0, & \text{otherwise} \end{cases} \quad (7-15)$$

The transfer matrix $P = \{p_{ij}\}$ of a social network G is defined as follows.

$$p_{ij} = \begin{cases} \frac{M(i, j)}{\sum_{v_k \in \text{outlink}[v_i]} M(i, k)}, & \text{outlink}[v_i] \neq 0 \\ M(i, j) = 0, & \text{otherwise} \end{cases} \quad (7-16)$$

Here, $\text{outlink}[v_i]$ represents the link-out nodes of node v_i . The transfer matrix P indicates that the node passes its authority equally to the nodes it links out to.

So, traditional PageRank mainly included two characteristics:

- (1) The node passes its authority equally to the nodes it links out to;
- (2) Each node is visited by other nodes at the same probability $1/n$.

The specific algorithm is as follows.

```

initialize ranks  $\pi$  0
while (not converged)
    for each vertex  $i$ 

```

$$\pi = \alpha P^T \pi + (1 - \alpha) \frac{1}{n} \mathbf{e}, \mathbf{e} = (1, 1, \dots, 1)^T$$

end

end

For example, the adjacency matrix town in Figure 7-1 is as follows.

$$M = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Then, the above adjacency matrix is converted into the transfer matrix.

$$P = \begin{bmatrix} 0 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1/4 & 0 & 0 & 1/4 & 0 & 0 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

Finally, repeated iteration is carried out by using the PageRank algorithm, to calculate the influence scores of each node(See Table 7-1).

Table 7-1 Influence scores

ID	PR	Inlink	Outlink
1	0.0250	v_5	v_2, v_3, v_6
2	0.0259	v_1	v_4, v_5
3	0.0562	v_1, v_7	v_4, v_6
4	0.4068	v_2, v_3, v_5, v_6, v_8	v_6
5	0.0298	v_2	v_1, v_4, v_7, v_8
6	0.3955	v_1, v_3, v_4	v_4
7	0.0357	v_5, v_8	v_3
8	0.0251	v_5	v_4, v_7


Then, the Personalized PageRank was proposed by Haveliwala Taher et al^[15] based on PageRank, as shown in the following equation. For example, element r_i indicates the degree of an individual's preference to the topic, the degree of novelty and sensitivity of the information released by an individual, etc.


$$\pi = \alpha P^T \pi + (1 - \alpha)r \quad (7-17)$$

In the Personalized PageRank, vector $\frac{1}{n}e$ is replaced by the personalized vector r .

Figure 7-11 shows the influence scores of two different types of personalized vectors in line with the case in Figure 7-1.

ID	π	r	ID	π	r	ID	π	r
1	0.0250	0.125	1	0.1024	0.65	1	0.0100	0.05
2	0.0259	0.125	2	0.0365	0.05	2	0.0103	0.05
3	0.0562	0.125	3	0.0515	0.05	3	0.0225	0.05
4	0.4068	0.125	4	0.3774	0.05	4	0.4384	0.05
5	0.0298	0.125	5	0.0230	0.05	5	0.0119	0.05
6	0.3955	0.125	6	0.3792	0.05	6	0.4825	0.65
7	0.0357	0.125	7	0.0177	0.05	7	0.0143	0.05
8	0.0251	0.125	8	0.0124	0.05	8	0.0100	0.05


 uniform vector


 bias on v_1



 bias on v_6

Figure 7-11 Personalized PageRank algorithm calculations

The traditional PageRank is only suitable for directed graph. For weighted directed graph^[11], the relevant scholars studied the weight PageRank algorithm. Figure 7-12 gave a weighted graph.

$$M = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix} \Rightarrow M = \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 2 & 0 & 0 & 3 & 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Adjacent Matrix
Weighted Adjacent Matrix

Figure 7-12 Weighted adjacency matrix

PageRank value of the weighted adjacency matrix is measured as follows.

Tunkelang et al. constructed an algorithm similar to PageRank targeted at the

followed-following relationship in Twitter, to measure the influence of individuals in Twitter, which uses the influences of fans to measure the influence of an individual; the higher the influence of the fans, and the smaller the number of other users being followed by these fans, the greater their contribution to the individual's influence. The model is constructed as follows:

ID	π	r
1	0.0250	0.125
2	0.0259	0.125
3	0.0562	0.125
4	0.4068	0.125
5	0.0298	0.125
6	0.3955	0.125
7	0.0357	0.125
8	0.0251	0.125

↑
Un-weighted

ID	π	r
1	0.0239	0.125
2	0.0255	0.125
3	0.0541	0.125
4	0.4142	0.125
5	0.0332	0.125
6	0.3902	0.125
7	0.0376	0.125
8	0.0213	0.125

↑
Weighted

Figure 7-13 Weighted PageRank values

(1) $\text{Influence}(X)$ = expected number of people who will read a microblog that X tweets, including all reposts of that microblog.

(2) If X is a member of Followers (Y), then there is a $1/\|\text{Following}(X)\|$ probability that X will read a microblog posted by Y , where $\text{Following}(X)$ is the set of people that X follows.

(3) If X reads a microblog from Y , there's a constant probability p that X will repost it.

From this model, it's easy to measure someone's influence recursively, assuming that we know the constant repost probability p :

$$\text{Influence}(X) = \sum_{Y \in \text{Followers}(X)} (1 + p \cdot \text{Influence}(Y)) / \|\text{Following}(Y)\| \quad (7-18)$$

7.3.3 Individual Influence Calculation Based on Behavior

In social networks, especially in Twitter, a user's influence is subject to its behavioral characteristics; the influence of a user is measured by four relationships: repost, reply, reintroduce (copy) and read^[16].

Literatures^[16] take into account a variety of network relationships in microblogs, in order to fully measure a user's influence in the topic level, as shown in Figure 7-14 (a), where the influence of user A by user B is represented in four types:

- (1) User A reposts the microblogs of user B by making use of an informal convention such as “RT @user” or “via @user”;
- (2) User A replies the microblogs of user B by making use of an informal convention such as “@user”;
- (3) User A reintroduces the microblogs which are similar to those previously posted by user B, but without acknowledgement of the source from user B;
- (4) User A read the message tweeted by user B.

Therefore, we can see that the influence network is a multi-relational network. Literature^[14] defined the following four types of networking relationships: Repost’s Network, Reply’s Network, Copy’s Network, and Read’s Network. For example, in Figure 7-14, the multi-relational influence network consists of three users and four relationships.

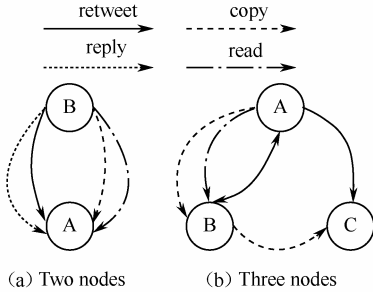


Figure 7-14 Multi-relation network

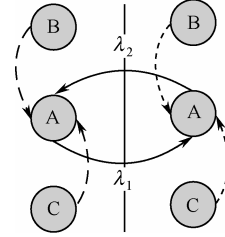


Figure 7-15 Inter-network and intra-network transition probability

The random walk process in the repost network graph $G_a^i = (V_a^i, E_{\text{Retweet}}^i, W(E_{\text{Retweet}}^i))$ is constructed as follows: user i reposts microblogs of his “friends” in the i -th topic space at certain transition probability under the influence of his friend. The random walk process in the repost network graph simulates the reposting behavior of users in Weibo. The transition matrix for the topic i , denoted as P_a^i , is defined as follows:

$$P_a^i(u_t^i | u_s^i) = \frac{w_a(u_s^i, u_t^i)}{\sum_{u' \in \text{out}(u_s^i)} w_a(u_s^i, u')}, \quad (7-19)$$

$w_a(u_s^i, u_t^i)$ is the frequency of user u_s^i reposting user u_t^i in the i -th topic space. $\sum_{u' \in \text{out}(u_s^i)} w_a(u_s^i, u')$ sums up the number of reposts by user u_s^i of the microblogs from all his

friends.

The random walk process in the reply network graph $G_b^i=(V_b^i, E_{\text{Reply}}^i, W(E_{\text{Reply}}^i))$ is constructed as follows: user i replies microblogs of his “friends” in the i -th topic space at certain transition probability under the influence of his friend. The random walk process in the reposting network graph simulates the reply behavior of users in microblog. The transition matrix for the topic i , denoted as P_b^i , is defined as follows:

$$P_b^i(u_t^i | u_s^i) = \frac{w_b(u_s^i, u_t^i)}{\sum_{u^i \in \text{out}(u_s^i)} w_b(u_s^i, u^i)} \quad (7-20)$$

$w_b(u_s^i, u_t^i)$ is the frequency of user u_s^i replying user u_t^i in the i -th topic space. $\sum_{u^i \in \text{out}(u_s^i)} w_b(u_s^i, u^i)$ sums up the number of replies by user u_s^i to microblogs from all his friends.

The “copy” behavior in microblog is somewhat a “repost” behavior, except that the “RT @B” or “via @B” tags are not explicitly used.

If two microblogs in a “copy” relationship are defined as a tuple $\langle p_t, p_s \rangle$, then all the microblog pairs in the “copy” relationship between the two friends is a binary set U . It can be deduced that the Copy Network graph $G_c^i=(V_c^i, E_{\text{Copy}}^i)$ is a weighted directed graph, and the weight $w_c(u_s^i, u_t^i)$ between user u_s^i and u_t^i is defined as follows:

$$w_c(u_s^i, u_t^i) = \sum_{\langle p_t^i, p_s^i \rangle \in U_{s,t}^i} \text{sim}(p_s^i, p_t^i) \times f(\Delta t_{p_s^i, p_t^i}) \quad (7-21)$$

$U_{s,t}^i$ is the binary set of the microblog pairs that are in the “copy” relationship between user u_s^i and u_t^i in the i -th topic space. The probability distributions of similarity and the time difference between the two microblogs are taken into account in the weight calculation. The higher the similarity between the two microblogs, the less the time difference between them, which indicates that the probability of “copy” is higher.

The random walk process in the copy network $G_c^i=(V_c^i, E_{\text{Copy}}^i, W(E_{\text{Copy}}^i))$ is constructed as follows: the user is influenced by his friend in the i -th topic space and will copy his friend’s microblog with a certain probability of transition. The random walk process in the copy network graph simulates the copy behavior of the user in Weibo. Let the transition probability matrix in the copy network in the i -th topic space be P_c^i , then the transition probability between the users is defined as follows.

In the copy network in the i -topic space, the transition probability of user u_s^i ’s random copy of user’s u_t^i microblogs is defined as

$$P_c^i(u_t^i | u_s^i) = \frac{w_c(u_s^i, u_t^i)}{\sum_{u^i \in \text{out}(u_s^i)} w_c(u_s^i, u^i)} \quad (7-22)$$

Where, $w_c(u_s^i, u_t^i)$ represents the weight of the “copy” relationship between user u_s^i and u_t^i ; $\sum_{u^i \in \text{out}(u_s^i)} w_c(u_s^i, u^i)$ represents the sum of the weights of user u_s^i 's “copy” relationship with all his friends in the i -th topic space.

The inference of the probability of a read relationship between users is related to the following three factors:

- (1) Users read with a higher probability the microblogs posted by friends who excel in quantity of posts
- (2) Users read with a higher probability the microblogs with a higher similarity of topics;
- (3) Users read with a higher probability the microblogs posted by friends with high similarity in the time series pattern of posting.

Therefore, the probability of user u_s reading the posts of his friend u_t is defined as follows:

$$P_{\text{read}}(u_s, u_t) = \frac{\tau_t \times \text{sim}(u_s, u_t) \times \text{simSeries}(u_s, u_t)}{\sum_{u \in \text{out}(u_s)} \tau_u \times \text{sim}(u_s, u) \times \text{simSeries}(u_s, u)} \quad (7-23)$$

Where, τ indicates the number of posts published by user u in the data set (not including the posts that are already in the repost, reply, and copy relationship). $\text{simSeries}(u_s, u_t)$ represents the time series similarity between user u_s and his friend u_t , and $\text{out}(u_s)$ represents the friends set followed by user u_s .

Users' influence from their friends is expressed as the random walk process in the four types of influence networks, which will also jump to another type of influence network at a certain probability. If the probabilities of a user staying in the repost network, reply network, copy network, and read network are respectively λ_1 , λ_2 , λ_3 , λ_4 and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$ is satisfied, then the user will jump from the repost network at a probability of $1 - \lambda_1$, from the reply network at a probability of $1 - \lambda_2$, from the copy network at a probability of $1 - \lambda_3$, and from the read network at a probability of $1 - \lambda_4$ to other networks.

According to the PageRank algorithm, the user not only travels along the network randomly, but also jumps to other nodes randomly at certain probabilities. Therefore,

considering the inter-node jump probability β and the inter-network jump probability λ , and let the transition probability matrix be B , the transition probabilities in the four networks in the i -th topic space is defined as follows.

(1) repost network:

$$\mathbf{B}_a^i(u_t^i | u_s^i) = \lambda_1 \times (1 - \beta) \times \frac{w_a(u_s^i, u_t^i)}{\sum_{u^i \in \text{out}(u_s^i)} w_a(u_s^i, u^i)} + \frac{\beta}{n} \quad (7-24)$$

(2) reply network:

$$\mathbf{B}_b^i(u_t^i | u_s^i) = \lambda_2 \times (1 - \beta) \times \frac{w_b(u_s^i, u_t^i)}{\sum_{u^i \in \text{out}(u_s^i)} w_b(u_s^i, u^i)} + \frac{\beta}{n} \quad (7-25)$$

(3) copy network:

$$\mathbf{B}_c^i(u_t^i | u_s^i) = \lambda_3 \times (1 - \beta) \times \frac{w_c(u_s^i, u_t^i)}{\sum_{u^i \in \text{out}(u_s^i)} w_c(u_s^i, u^i)} + \frac{\beta}{n} \quad (7-26)$$

(4) read network:

$$\mathbf{B}_d^i(u_t^i | u_s^i) = \lambda_4 \times (1 - \beta) \times \frac{w_d(u_s^i, u_t^i)}{\sum_{u^i \in \text{out}(u_s^i)} w_d(u_s^i, u^i)} + \frac{\beta}{n} \quad (7-27)$$

Let $r^i(u)$ be user u 's ranking in the i -th topic space; taking into account the random walk of the user in the four kinds of network, user u 's ranking in the i -th topic space is defined as follows:

$$\begin{aligned} r^i(u) = & \sum_{(u_t^i, u) \in E_{\text{Retweet}}^i} \mathbf{B}_a^i(u | u_t^i) r^i(u_t^i) + \sum_{(u_t^i, u) \in E_{\text{Reply}}^i} \mathbf{B}_b^i(u | u_t^i) r^i(u_t^i) \\ & + \sum_{(u_t^i, u) \in E_{\text{Copy}}^i} \mathbf{B}_c^i(u | u_t^i) r^i(u_t^i) + \sum_{(u_t^i, u) \in E_{\text{Read}}^i} \mathbf{B}_d^i(u | u_t^i) r^i(u_t^i) \end{aligned} \quad (7-28)$$

That is, the ranking of a user is mainly determined by the probability of its followers' random jump to the user.

7.3.4 Individual Influence Calculation Based on Topics

In social networks, the influence of an individual varies with different topics. Weng Jianshu et al^[17] measured the influence of an individual on each topic taking into account both the topical similarity between users and the link structure in the Twitter data set. Given a topic t , each element of matrix P_t , i.e. the transition probability of the random surfer

from follower s_i to friend s_j , is defined as:

$$P_t(i, j) = \frac{|T_j|}{\sum_{a: s_i \text{ follows } s_a} |T_a|} \times \text{sim}_t(i, j) \quad (7-29)$$

$|T_j|$ is number of tweets published by s_j , and $\sum_{a: s_i \text{ follows } s_a} |T_a|$ sums up the number of tweets published by all of s_i 's friends.

Topical difference between two twitterers s_i and s_j can be calculated as: $\text{dist}(i, j) = \sqrt{2D_{\text{JS}}(i, j)}$.

$D_{\text{JS}}(i, j)$ is the Jensen-Shannon Divergence^[18] between the two probability distributions DT'_i and DT'_j , which is defined as:

$$D_{\text{JS}}(i, j) = \frac{1}{2} (D_{\text{KL}}(D'_i \parallel M) + D_{\text{KL}}(D'_j \parallel M)) \quad (7-30)$$

M is the average of the two probability distributions, i.e. $M = \frac{1}{2}(D'_i + D'_j)$. D_{KL} is the Kullback-Leibler Divergence which defines the divergence from distributions Q to P as:

$$D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

So, the similarity of two topics is defined as follows.

$$\text{sim}_t(i, j) = 1 - |DT'_i - DT'_j| \quad (7-31)$$

Figure 7-16 gave an example of TwitterRank.

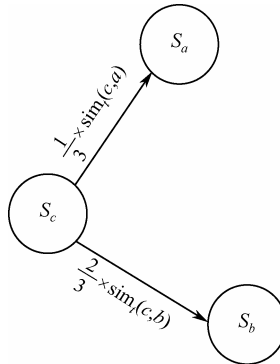


Figure 7-16 Schematic diagram of TwitterRank diagram

7.4 Summary

With the rapid development of online social networks and the rapid growth of online users, social networks for the purpose of making friends, information sharing and so on have become an ideal platform for people to share information, sell goods, express opinions and develop personal influences. The influence analysis and modeling of online social networks constitutes an important part of social network analysis. Analyzing the influence modes and the ways of influence diffusion will not only deepen our understanding of people's social behaviors from a sociological point of view, but also provide a theoretical basis for public decision-making and public opinion guidance. At the same time, it can also promote the communications and disseminations of activities in political, economic, cultural and other fields, which bears important social significance and application value.

Although fruitful results have been achieved in the field of influence analysis in social networks, we believe that at least the following questions require further study and exploration:

(1) Due to the large number of social networking users, and the very complex relationships between users, qualitative analysis of social influence in such an environment is also subject to many factors and interference. Although many researches attempted to objectively and accurately clarify the relationship between influence and other factors, they still cannot effectively solve this problem. Such a situation is associated with both the complicated generation and diffusion mechanisms of social influence, and with the definition of influence itself. The existing concept of influence is nothing more than a description of the effect of influence, which, in essence, does not explain the problem as to "what is influence", resulting in the enormous social influence models, with the lack of benchmark comparison models and methods. Perhaps we cannot precisely define the concept of social influence, but as far as the specific environment of online social network is concerned, it is necessary to study indicators of evaluating social influence, to provide directional guidance for the design of new models, so that they can describe the complexities of online social networks more accurately.

(2) At present, the social influence modeling methods can be divided into two categories: one is the empirical method and the other is the inference method. The empirical method summarizes mathematical models that comply with samples based on the observation and analysis of experimental data, and then achieve

parameter values in the models by fitting with the actual data. The inference method directly derives influence models according to the relevant theory, and also determines parameter values in the models by means of learning and fitting. Both methods have their own advantages and successful applications; however, neither can universally and accurately depict the influence in social networks. To change such a situation, it is required to seek breakthroughs in theoretical work such as the definition of influence, the relationship between social information and influence, and so on; it is also necessary to seek improvements in the modeling methods. We can make use of such information as social network topology, user interaction data and action records, to analyse the generation and dissemination processes of user influence in social networks globally from multiple perspectives; on occasions of higher real-time requirements, we can also consider using incremental models to reduce the amount of computation.

References

- [1] Elihu Katz. Personal influence: the part played by people in the flow of mass communications. Glencoe, Illinois, 1955.
- [2] Everett Rogers. Diffusion of innovations, Simon and Schuster, 1962.
- [3] Noah Friedkin. A structural theory of social influence, Cambridge University Press, 2006.
- [4] Mark Granovetter. The strength of weak ties[J]. American journal of sociology, 1973, 78(6): 1.
- [5] Linton Freeman. A set of measures of centrality based on betweenness. Sociometry, 1977: 35-41.
- [6] Akshay Java; Pranam Kolari, et al. Modeling the spread of influence on the blogosphere. Proceedings of the 15th international world wide web conference, 2006.
- [7] Goya Amit and Bonchi Francesco, et al. Learning influence probabilities in social networks. Proceedings of the third ACM international conference on Web search and data mining, ACM, 2010.
- [8] Tang Jie and Sun Jimeng, et al. Social influence analysis in large-scale networks. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009.
- [9] Liu Lu and Tang Jie, et al. Mining topic-level influence in heterogeneous networks.

- Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010.
- [10] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 1999, 46(5): 604-632.
 - [11] Larry Page and Brin Sergey, et al. The PageRank citation ranking: Bringing order to the web, 1999.
 - [12] Cha Meeyoung and Haddadi Hamed, et al. Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM*, 2010, 10: 10-17.
 - [13] Brandes, Ulrik. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 2001, 25(2): 163-177.
 - [14] Romero Daniel and Galuba Wojciech, et al. Influence and passivity in social media. *Machine learning and knowledge discovery in databases*, Springer, 2011: 18-33.
 - [15] Haveliwala Taher. Kamvar Sepandar, et al. An analytical comparison of approaches to personalizing PageRank, 2003.
 - [16] Ding Zhaoyun, Jia Yan, Zhou Bin, Han Yi. Mining topical influencers based on the multi-relational network in micro-blogging sites[J]. *China Communications*, 2013, 10(1): 93-104.
 - [17] Weng Jianshu and Lim Ee P, et al. TwitterRank: Finding topic-sensitive influential twitterers [C]. In the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10). New York, USA, 2010: 261-270.
 - [18] Bent Fuglede (2004). Jensen-Shannon divergence and Hilbert space embedding. *International Symposium on Information Theory*, 2004. ISIT 2004. Proceedings. p. 30. doi:10.1109/ISIT.2004.1365067.

Collective Aggregation and the Influence Mechanisms

There is no doubt that collective power plays a key role in the occurrence of every economic phenomenon, political decision or social transformation. One of the best sellers published in 1897, *The Crowd* ^[1] (Gustave Le Bon, 1841—1931), even further provoked people's interest in seeking an in-depth understanding of groups, especially group behavior. As the great master of media Marshall McLuhan (1911—1980) commented, media is a hidden factor that shapes our history and society, and because online social networks enable numerous individuals in reality to easily gather through the Internet, they have become a new force in promoting social evolution ^[2]. Collective intelligence and group polarization^① are the two extreme ways in which society is influenced by collective aggregation. On the one hand, online social networks allow users to contribute their ideas to the shaping of collective intelligence; on the other hand, they lead to a loss of users' self-judgment during group interactions, resulting in the tendency of identical irrationality in group behaviors.

The aim of this chapter is to explore collective aggregation and the mechanisms that influence it from the perspectives of collective intelligence and group polarization. The content of the chapter has practical significance regarding how collective intelligence is formed through online social networks and how the advantages of group polarization can be exploited in online social networks while avoiding its disadvantages.

The chapter is organized as follows: Section 8.1 defines the concept of a group and the concept of a group in a social network from different perspectives and provides unified

① Briefly, collective intelligence is a group behavior in which many individuals generate problem-solving ability superior to individuals through mechanisms such as competition and cooperation, differentiation and integration, and feedback and selection. Group polarization refers to the effect of discussions among group members on individual members' opinions or decisions in group decision-making situations, which leads to behavioral consistency within groups.

definitions that are used in this book. Sections 8.2 and 8.3 expound upon the mechanisms through which collective intelligence and group polarization are respectively generated. First, the developments in current researches are discussed based on the concepts and relevant theories, the generation conditions and the major influencing factors of collective intelligence and group polarization; secondly, the analytical models for the two types of group behaviors are further explained; finally, Section 8.4 provides a brief summary of this chapter.

8.1 Introduction

A group is a cross-disciplinary concept, and researchers in many disciplines have defined it from different perspectives.

Definition 8-1:

David W. McMillan (1917—1998), a well-known professor of psychology at Harvard University, defined a group as a feeling^[3]; that is, the connection, membership, and shared beliefs among group members, where members meet their needs through the form of organization. A group should contain four elements: membership, influence, integration and fulfillment of needs and shared emotional connection. From the perspective of psychology, a group should provide its members with a sense of belonging; individuals are able to influence the decision-making process of their group, and their needs can be satisfied in the group. This definition also emphasizes the common emotional ties within a group; namely, members share the same history, location, hobbies and so on, which compose the boundary that separates a group from outside. Psychological researchers mainly analyze the psychological feelings of individuals within a group and thus define a group from the perspective of individual psychology.

Definition 8-2:

In the field of sociology, the concept of a group is also a hot topic. In 1981, Marvin E. Shaw^[4] defined a group as “two or more persons who are interacting with one another in such a manner that each person influences and is influenced by each other person”, where there must be interactions among members within the group. He also believed that the distinction between “a group” and “a crowd” lies in the influence among the members of the group, and because of such phenomena as “group facilitation” and “group polarization”, a group is not simply equal to the aggregation of individuals in terms of function. The sociological definition also includes the mutual influence among members of a group; furthermore, from the point of view of group function, it emphasizes that a group is not the aggregation of individuals; the existence of such phenomena as group polarization

increases the uncertainty of group decisions and performance.

Definition 8-3:

Anthropologist Marcello Andrea Canuto ^[5] defined a group as a collection of individuals geographically adjacent to one another, who have regular contact and interactions. The definition of a group from the field of anthropology is similar to that from the field of sociology: they both indicate the presence of interaction among individuals in the group. However, the anthropological definition of a group emphasizes the geographical proximity of the members and their interaction in reality with other group members.

The definition of a group in this book is comprehensive, as it is based on the definitions from psychology, sociology and anthropology, with an emphasis on information sharing, mutual influence, aggregation and other characteristics of individuals in a group. Therefore, in this book, a “group” is defined as a number of people who group together, either spontaneously or organizationally, by sharing information and working for common goals, where each member influences and is influenced by each other member.

A group in an online social network refers to a number of people who motivated by common goals or interests, group together through online social networks, either spontaneously or organizationally, for the purpose of sharing information, where each member influences and is influenced by each other. It has the following characteristics:

- (1) Relying on social networks as the carrier.
- (2) With exchanging information as the purpose.
- (3) Free of geographical restrictions.

Members of a group in a social network influence one another by sharing information and interacting with other members through online social networks; therefore, a group in a social network is a virtual community. In 1998, Iain Pears noted that although some online virtual communities are based on geographical distributions, most of them have no geographical restrictions, which is a distinct difference from the traditional concept of a group ^[6].

Because people can link with each other and share information on the Internet, thousands of millions of independent Internet users are connecting with each other and creating numerous lively virtual communities with unified behaviors. Consequently, the Web 2.0 world is full of countless virtual groups with different sizes, various purposes, and dynamic variations. Information flows in these virtual communities, and the viewpoints contained in such information become a force driving individuals to quickly gather to form

a virtual community. Hence, this chapter explores the phenomenon and the influence mechanism of collective aggregation in social networks from the perspective of information sharing.

Online social network, as a public platform where individuals come together and form group opinions, has undoubtedly become an effective catalyst for collective intelligence and group polarization due to its high levels of interaction, aggregation and uncertainty. This chapter focuses on collective intelligence and group polarization in social networks and expounds the concepts, theoretical basis, generation, conditions, influence factors and analytical models of these two topics. It then anticipates the prospects of this specific field of research in development.

8.2 Mechanisms Engendering Collective Intelligence

8.2.1 Collective Intelligence

Collective intelligence emerges from cooperation and competition among a number of individuals. We may say that research on collective intelligence is a long-term proposition in the field of collective behaviors, which covers the hierarchy from quark to bacteria, to plant, to animal and to human society.

The concept of collective intelligence originated from biology, initially from the observations of entomologist William Morton Wheeler (1902—1981) ^[7]. In 1911, William Morton Wheeler found that ants behaved like an animal cell and that they seemed to have a collective mind, which he referred to as a larger creature; namely, the aggregation of ants seems to create a “super-organism”. In 2010, Howard Bloom depicted the evolution of swarm intelligence in his book and noted that swarm intelligence started to play a role from the origin of life. Notably, this biological definition emphasizes that the aggregation of individuals makes the group more powerful.

In the field of sociology, Émile Durkheim (1859—1917) considered that a society group has higher intelligence than the individuals it contains in terms of time and space ^[8]. George Pólya, a pioneer in collective intelligence research, defined collective intelligence as “the capacity of human communities to evolve towards higher order complexity and harmony, through such innovation mechanisms as differentiation and intergration, competition and collaboration”. Sociologists borrowed the concept of swarm intelligence from biology,

applied it to human society, and called it collective intelligence. However, this sociological definition describes collective intelligence as a capacity of human communities, rather than a biological phenomenon, which emphasizes that collective intelligence is derived from the initiative and motility of group members.

To understand this connotation of collective intelligence, we have to describe the basic features of collective intelligence systems. Martijn C. Schut (2010) concluded that previous studies have summarized five characteristics of a collective intelligence system^[9] :

(1) Integrity and locality. If a collective intelligence system is divided into two levels, the individual level comprises individuals composing the system, which is the locality feature of the system; the overall level takes the entire system as a whole, which is the integrity feature of the system.

(2) Randomness. A complex system typically has some random characteristics, which allows the system to enter a self-organized critical state. Such a critical state enables the system to be at the edge of chaos: disordered on the one hand, but structured and ordered on the other hand.

(3) Emergence. The simplest description of emergence is that “the whole is greater than the sum of its parts”, which is especially the case in a collective intelligence system.

(4) Redundancy. This characteristic refers to the fact that the same knowledge may be embodied in a series of different positions in a system, and this knowledge can be either concrete knowledge, such as interaction rules, or information per se. That is to say, a large number of individuals may follow the same rules if we view it from a simple dimension; however, such rules or information may not apply to all individuals within the system if we examine it from a more complex perspective.

(5) Robustness. Similar to redundancy, the robustness of a system can resist system failures. If one rule appears more than once in a system, even if an individual misses the rule, other individuals in the system can also follow the rule.

Based on the definitions and theories above, we define collective intelligence in this book as the capacity of a group of individuals to fulfill tasks which separate individuals may have difficulties to conquer through such innovation mechanisms as competition and cooperation, differentiation and integration, and feedback and selection.

Online social networks provide a more efficient platform through which collective intelligence is realized. In this new era, new technologies on which collective intelligence relies have played a role in increasing the efficiency of knowledge sharing on the Internet,

which has greatly speeded up the flow of global culture and knowledge. Further, collective intelligence in social networks not only increases the amount of information but also plays an important role in improving and maintaining the quality of such information.

8.2.2 Self-determination Theory and Collective Intelligence

Self-determination theory, a theory about the motivational process of human self-determination, was proposed by an American psychologist Edward Deci in the 1980s. The theory suggests that self-determination potentially drives people to choose the behaviors that are both line with their own personal interest and conducive to their individual developments; thus, self-determination is the main component of intrinsic motivation behind human behavior^[10]. According to the self-determination theory, human behavior is mainly driven by three needs: autonomy, competence, and interpersonal relatedness.

The need for autonomy is the core driving force of human behavior. In brief, an individual is the physical abstraction of an actual system or the functional unit of the system. Individuals can take certain autonomous actions to meet the designed goals in a certain environment, and they can perceive the environment and adapt themselves to changes in the environment. Since their advent, social networks have become the infrastructure for the self-organization of individuals, where individuals can independently select or post contents and show more independence and autonomy. Therefore, the need for autonomy is more intense in social networks.

Competence refers to the underlying personal characteristic that can distinguish outstanding achievers from mediocre individuals in a certain workplace, organization or culture. It can be the motivation, character, self-image, attitudes, values, expertise, cognitive skills, behavioral skills, or any other qualities of a person that can be reliably measured or quantified, and that can serve to significantly distinguish excellence from average. Although individuals in social networks are equal, competence can differentiate them according to their performances of content contribution. For instance, opinion leaders and forum hosts may be distinguished from ordinary Internet users. However, as individuals always attempt to distinguish themselves from others through certain performance, the need for competence still exists among Internet users.

In sociology, interpersonal relatedness is defined as a kind of social relationship that is established in the process of production and living activities in society, whereas psychologists

define it as the direct psychological connection between people established in the process of communications. In daily life, interpersonal relatedness often refers to general interactions between people, including those in familial relationships, social relationships, alumni (classmate) relationships, teacher-student relationships, and employment relationships, such as those between colleagues and supervisors. In the world of social networks, people do not communicate face to face any longer. Although the need for interpersonal relatedness may be diminished in such a context, the need for a sense of belonging still induces people to seek their own group on the Internet, which is the reason why online communities, forums and interest groups exist. Moreover, the need for a sense of belonging is associated with intrinsic motivation. Self-determination theory assumes that interpersonal relatedness engenders a kind of dynamic force. When an individual has a sense of security and belonging in an environment, he or she will have greater intrinsic motivation. Therefore, the need for interpersonal relatedness is still very important on the Internet.

With self-determination theory, we can better analyze the intrinsic motivation behind users' behavior of contributing to collective intelligence, and quantify the degree of their contributions. In contributing to collective intelligence, individuals' needs for self-determination and competence are fully met, but it may not be the case in a traditional production organization. In online communities, there is no hierarchy or task allocation; participants choose tasks entirely based on their own interests and needs. Meanwhile, collective intelligence is a process through which participants work out a solution to a problem by gathering wisdom from all the participants, and this process constantly relies on cooperation among the participants. The participants can discuss problems through online communication, interactive discussion and so forth, which further promotes interaction and communication between participants, fosters the motivation for exchanges and satisfies participants' need for interpersonal relatedness.

In the Web 2.0 environment, the greater openness and decentralized environment of social platforms allows participants to obtain their desired information and publish their views regardless of time and geographical restrictions. The impetus to aggregation in a network community is the interest shared by participants' shared interest, which increases the participants' motivation for and level of participation.

Therefore, social networks sufficiently meet participants' needs for self-determination, competence and interpersonal relatedness, providing a favorable environment for the generation of collective intelligence.

8.2.3 Conditions Engendering Collective Intelligence

In May 1968, an American nuclear submarine called “Scorpion” suddenly disappeared in the North Atlantic, and no one could determine where it sank. The U.S. Navy searched for several months, to no avail. At that time, a man named Craven organized experts, including mathematicians, submarine experts and fishing specialists, to analyze the possible location of the submarine. Instead of organizing a group discussion, which is the conventional way, he required them to analyze the possible location back to back, and then performed a comprehensive analysis of the preliminary results. The final results turned out to be different from any of the individual results. Surprisingly, the submarine was found 22 yards away from the final analysis results.

This story is from the book “The Wisdom of Crowds”^[11] by James Surowiecki. In his work, he stated that under the right circumstances, collective wisdom could be very surprising; it is generally superior to the judgment made by the person with the highest IQ alone. However, on no account can we ignore Tom Hayes’s explanation (2010) about the prerequisites for this sentence—“the right condition”: independence, diversification and dispersion^[12]. According to James Surowiecki, “if we can gather together people from different industries, different areas, with different knowledge backgrounds, they have the ability to make the right decisions on major issues is superior to a couple of so-called elites.

In 2006, Don Tapscott and Anthony D. Williams summarized three guidelines to produce extensive group collaboration and ultimately engender powerful collective intelligence: ①foster a goal of collaboration, ②guide individuals to contribute efforts independently, and ③integrate the independent contributions^[13]. Tom Hayes (2010) further showed that if a group has a very high degree of homogeneity, their output will be extremely lack of diversity^[12]. Bede Miller also noted that only when all members make decisions independently and bear responsibility for their actions can they collectively light up the light of wisdom. If members follow each other blindly, the group will produce foolish outcomes. In 2010, Thomas W. Malone stated that for a wise collective, regardless of whether it is a collective of ants or lawyers, its members must be independent and bear responsibility^[14].

This book argues that there are four necessary conditions for the generation of collective intelligence:

First, there must be specific goals for collaboration.

Secondly, individuals in a group must have the ability to think independently; that is, they must have independence. Indeed, the independence of individuals is a more important factor in collective intelligence than intelligence at the individual level, as it ensures that each individual will not be negatively affected by others, that they can think independently, tap into and stimulate their maximum potentials, and that they can contribute more accurate and creative ideas.

Thirdly, individuals within a group must be diverse, heterogeneous and dispersed. Diversity of ideas is a prerequisite for ensuring adequate information. Thus, the group should try to attract people with different knowledge backgrounds and rich experience and guide and encourage each person to form unique, complementary insights that can be integrated to stimulate greater collective potentials. Dispersion and decentralization allows full respect for individual differences and personality, so that every user's opinion can get the attention of the group. By exploiting users' unique perspectives, the group can establish a more comprehensive and accurate understanding of the problem.

Fourthly, a mechanism is required to eventually concentrate the decentralized intelligence, so as to collect the individual views, and create the group-level decision-making intelligence. The collection mechanism is not the simple adding or averaging, but rather the comprehensive consideration of various factors. It is make the optimal decision made on the basis of listening to the views of the members.

It is true that a group can produce intelligence, but it does not always produce collective intelligence of high quality. One of the factors affecting the quality of collective intelligence is the size of the group. Furthermore, individuals' speculation and judgement may comprise both real information and errors. James Surowiecki noted that if the error can be stripped away, then the remainder is the real information. If a sufficiently large group comprising individuals from various industries who are not influenced by one another, makes a forecast or estimation with respect to a problem and if the result is then averaged, the error made by each individual will be offset by the mean value. However, it is worth noting that sufficient information is a prerequisite for offsetting the error. For example, if a group of children without any experience in the stock market are engaged to predict the market trends, the result will most likely be erroneous even if all the above conditions are met.

8.2.4 Factors Influencing Group Intelligence

How much intelligence a group is able to produces is quite uncertain, for it is subject

to many factors. These factors include the size of the group, heterogeneity of the group members, the systems of participation and withdrawal of the group members, communication channels between the group members, mode of organizational management, the special role of conflicts and so on.

1. Group Size

In 2012, Sinan Aral and Dylan Walker^[15] demonstrated that group size is one of the most important factors that affect group intelligence and that groups with a limited size may not generate group intelligence.

Some scholars believe that group size is positively related to group intelligence. For instance, German sociologist Georg Simmel (Georg Simmel, 1858—1918)^[16] argued that small groups were not able to achieve scientific group decision making and that a group's ability to solve problems increases with the size of group. Stenfan Krause and Brent Gallupe's experiment also showed that larger group size leads to group decision making of higher quality. Moreover, Ioanna Lykourantzou^[17] and Wu-Chih Hu^[18] believed that group size is an important factor in group collaboration and large-scale social networking activities and that larger groups can attract more members and lead to greater collaboration.

However, neither Western nor domestic researchers have quantitated the specific sizes of groups. According to the definition of group size in social statistics, groups of large size should contain no fewer than 2000 individuals. Furthermore, group size should be quantified in units of one hundred or one hundred thousand if information exchange, communication and statistical techniques are taken into account.

2. Heterogeneity of Group Members

Heterogeneity is a positive factor for group intelligence. A famous biological behavior scientist Thomas D. Seeley^[19] believed that, a diversity of views is of great value for groups, and that groups with more diverse characteristics are more adept at solving problems.

Thomas D. Seeley^[19] wrote in his book *The Wisdom of the Hive* that diversity in a swarm of bees allows the swarm to find nectar sources more efficiently and that such efficiency depends on the colony-specific division of labor in the hive and the collaboration of bee species. Scott E. Page^[20] held the opinion that diversity in the view of group members is very valuable for helping groups solve problems. Everett Stiles^[21] found that the process of achieving group intelligence is the process of collective creation conducted by a group of individuals with different motives and goals. Indeed, diversity

in a group can guarantee the independence of ideas, increase the number of creative ideas, and enrich the types of creative ideas.

While group member heterogeneity has a positive effect on group intelligence, it is worth noting that the diversity of group members still relies on the amount of information which each individual possesses; otherwise, the diversity of the group members will have no effect on group intelligence.

3. Participation and Withdrawal of Group Members

In addition to studying the characteristics of the main body of a group, De Liddo Anna et al. ^[22] also found that the quality of group intelligence is also closely related to the participation system of a group. De Liddo Anna ^[23] stated that group intelligence can be roughly divided into two forms: unconscious and conscious. In social media networks, the views, ratings, reviews, purchase records, and so forth of users and the social networks established therefrom constitute the unconscious group intelligence. By contrast, group intelligence consciously deployed and produced by a group which hold the goal of collaboration is the conscious group intelligence. Comparatively, the quality and level of unconscious group intelligence is lower than that of conscious group intelligence. From another perspective, when solving problems such as urban planning and climate change, conscious participation is essential for achieving high-level group intelligence through high-level awareness. Therefore, the participation system of group members has certain effect on group intelligence (Dai Yang, 2014).

Due to the effect of “The Spiral Of Silence” in the knowledge economy ^[24], the lowered threshold for withdrawal from groups makes it possible for individuals holding objections to withdraw or be removed from the group, which accelerates the polarization of the group and has a serious negative effect on group intelligence. Since the proposition of the threshold theory by Granovetter, researches on the relationship between withdraw mechanisms and group intelligence have flourished. Given a lower threshold for withdrawal, rational individuals are more likely to withdraw from a group; the members who remain in the group are those who have a higher threshold for withdrawal. In such a case, the homogeneity of the group increases, while the diversity decreases. As a result, the withdrawal of group members exerts an influence on group intelligence.

4. Communication Channels Between Group Members

In 1996, Mohamed A. Amin et al. suggested that modern information and communication

technologies would enable instant information exchange among group members, thus promote the generation and communication of groupthink, which has a positive impact on group intelligence ^[25].

Interaction and information sharing among individuals is the key to achieve group intelligence. The popularity of the Internet, especially the generation and application of Web 2.0 technology, has injected new vitality into group intelligence. Indeed, such convenient technologies allow users to instantaneously interact through the Internet, without programming skills, and enable them to directly participate in sharing and creating contents. As stated by John Smith in 2005: the extensive applications of the Internet and communication technology in human production and life once again stimulated people's awareness of group intelligence. In 1999, Pierre Levy also stressed that network technology had brought some human-friendly effect for human beings. Such effect arises from the collision of all kinds of views on the Internet, which engenders group intelligence and verifies the contribution of each individual.

More and more scholars have suggested that the formation of group intelligence is a dynamic interactive multi-stage process and that information exchange and communication technologies play a pivotal role in this process. Thus, new interaction channels between the individuals of a group will have a positive impact on group intelligence.

5. Mode of Organizational Management

The mode of organizational management plays an extremely important role in producing group cohesion. Cartwright and Dorwin ^[26] stated that group cohesion can promote group members to act for the sake of the whole group to the greatest extent possible and thus foster autonomy of individuals. The impact on group intelligence of group cohesion mainly depends on its influence on group decision-making. However, improper manifestations of group cohesion will have an adverse impact on group intelligence. When other manifestations of groupthink (such as command-style leadership) arise, too much cohesion will lead to degraded group decision making ^[27]. For example, centralized organizational management can promote the phenomenon of group polarization, shorten the decision time and thus negatively affect group intelligence ^[28].

As the above studies have shown, the mode of organizational management has certain influence on group intelligence.

6. Special Role of Conflicts

Rather than affecting the quality of group decisions, appropriate conflicts play a

positive role in forming group intelligence^[29] in most cases.

Liu (2011) suggested that to prevent groupthink from restricting the decision-making process, a reasonable conflict must be created in the group. This type of conflicts does not refer to interpersonal conflicts; rather, it is a “brainstorming” method designed to avoid generating the bandwagon effect during the process of group decision making. If conflict occurs in a harmonious atmosphere, and if a decision is made by people from different professional backgrounds and with different thinking styles, there will be a higher level of participation and cooperation in discussions, and they will be more likely to ultimately arrive at a good decision. This approach fosters clear-mindedness in group decision making and prevents the decision-making process from becoming irrational. Therefore, under certain circumstances, conflict can have a positive effect on the group intelligence.

8.2.5 Analytical Models of Collective Intelligence

1. Model of Collective Intelligence Based on Bayesian Theory

In real life, we are forced to make decisions while facing risks and uncertainties. The horse racing problem provides a good example. Gamblers often choose their horses based on their judgment of the possibility of each horse winning the game. A betting market, such as that for a sporting event like horse racing, provides a typical scenario where the behaviors of individuals aggregate into collective intelligence. It should be noted that investors are often attracted by investments with high profitability.

Consider an event in which ten horses participate in a horse racing as a simple example. If this is the only information available for us to make a wager, then we can do nothing but selecting a horse randomly; the possibility of winning in such a case is only 0.10. Such an approach will certainly lead to losing the bet. However, this is a problem that can be solved with Bayesian logic^[30]. In fact, as each horse has participated in this event several times, they all have a racing history. If horse No. 1 is sure to win whenever it participates in a game, while horse No. 2 is sure to lose whenever it participates in a game, then we have a real and well-documented basis for a bet: wager money on horse No. 1 rather than on horse No. 2. All such information helps us to make a better prediction on the winning horse than simply choosing one from the ten horses randomly. The process of analyzing these factors is the Bayesian process.

The modeling based on Bayesian theory is probabilistic reasoning. That is, it is a

method to perform reasoning and decision-making tasks when all conditions, except the probabilities of occurrence, are uncertain. Take the stock forecasting for instance: we calculate the posterior probability of the same stock price by using a Bayesian estimation formula with different observations in different periods on the basis of knowledge of prior information of the stock price; then, we can make a reasonable estimation of the trends in the fluctuation of the stock price.

Taking stock price forecasting as an example, we introduce these basic procedures for the research of collective intelligence based on Bayesian theory, which comprise the following steps.

(1) Build a research model on the basis of a reference review of relevant research results on collective intelligence and Bayesian theory. Bayesian predictive models are established mainly based on the judgment of prior information and sample based informations. Prior information refers to the non-sample information based on experience and historical data; it is a random variable with respect to the required unknown parameter θ . By contrast, sample information refers to the information about the unknown parameter θ which is further obtained from a sample taken from the total. To make forecasts, the prior information is combined with the posterior information to obtain the posterior distribution of parameter θ .

For example, Yuqiu Sun and Shengtao Chen (2003)^① proposed a model based on the Bayesian decision making method to predict stock prices. They divided the stock prices into k intervals, denoted as $E_1 \dots E_k$, and the probability that a predicted object would fall into the interval E_i is denoted as p_i . They treated the event that the stock prices fall into a certain interval as an n -fold Bernoulli experiment and assumed that the prior distribution of p_i obeys an incomplete β distribution. They then acquired the density function of p_i , and were able to obtain the sample density function because the conditional distribution of the sample with respect to parameter p_i obeys a binomial distribution^[31].

Combining both methods, we are able to obtain the joint distribution of the samples and parameters (predicted value) and therefore determine the posterior distribution of p_i . Further, according to Bayesian estimation theory, in the condition of square loss, the Bayesian estimation of p_i becomes the mean of the posterior distribution, and the final Bayesian estimation formula can then be obtained.

(2) Obtain the quantitative data required to validate the model, based on the results of

① For more information, refer to the following paper: Yuqiu Sun, Shengtao Chen, Bayes decision method applied in the stock price forecasting [J]. Guangdong Polytechnic Normal University, 2003 (4): 78-80.

previous researches regarding the issue of collective intelligence. By establishing a model, we are able to obtain the Bayesian estimation formula. However, to carry out specific forecasts, we have to collect real sample data. For example, in the model of Yuqiu Sun and Shengtao Chen (2003) for forecasting stock prices, they obtained through data collection the closing price of stock No. 000029 in the Shenzhen stock market from February 1, 2002 to April 5, 2002, where the values were 5.38, 5.60, 5.59, 5.28, 5.42, 5.41, 5.95, 6.21, 6.33, 6.26, 6.64, 6.86, 7.55, 8.31, 9.14, 9.76, 9.64, 9.92, 9.74, 10.30, 9.66, 10.63, 11.00, 10.92, 10.77, 10.24, 10.32, 10.78, 10.99, 10.55, 10.31, 10.73, 11.07, 10.51, 10.41, and 10.35.

(3) Conduct statistic classification of the collected data and carry out further processing and computation. Pugging the sample data into the Bayesian formula, we will be able to determine the probability distribution of the predicted values obtained after the collection of prior information and sample information. First, Yuqiu Sun and Shengtao Chen (2003) classified and analyzed the statistical information in the original data and then initially divided it into different groups based on their states. In the following example, the data is divided into seven groups and the number of observations falling in each group is counted. Then, the multilayer prior density of each state is calculated. The results from the above steps are shown in Table 8-1.

Table 8-1 Outcomes of the Bayesian model calculation

i	E_i	Range of states	Number of observations under state E_i	\hat{P}_i				
				$C=2$	$C=3$	$C=4$	$C=5$	$C=6$
1	E_1	<6.00	7	0.14909	0.14856	0.14803	0.14752	0.14702
2	E_2	[6.00,7.00)	5	0.10144	0.10110	0.10077	0.10045	0.10012
3	E_3	[7.00,8.00)	1	0.01520	0.01517	0.01515	0.01512	0.01509
4	E_4	[8.00,9.00)	1	0.01520	0.01517	0.01515	0.01512	0.01509
5	E_5	[9.00,10.00)	6	0.12504	0.12461	0.12418	0.12376	0.12335
6	E_6	[10.00,11.00)	14	0.32589	0.32459	0.32337	0.32250	0.32124
7	E_7	≥ 11.00	2	0.03477	0.03468	0.03460	0.03451	0.03443

(4) Draw conclusions based on the statistical results. As the results of Yuqiu Sun and Shengtao Chen (2003) show, to analyze the distribution of predict values, we can judge their future states by finding the maximum probability with which they fall into an interval. The results of the analysis by Yuqiu Sun and Shengtao Chen (2003) predicated that the stock price on April 8, 2002 would be in the interval of [10.00,11.00), and indeed, the price of the stock on April 8, 2002 was 10.35, indicating that the prediction is accurate. As can be seen from the above examples, it is feasible to use a Bayesian model to predict stock prices;

moreover, this method can not only predict stock price trends but also accurately forecast the scope of the fluctuation of stock prices.

2. Collective Intelligence Model Based on Ant Colony Algorithm

Ants are high social creatures, which live in communities and forage for food collectively. In the process of finding food, each ant forages alone but “remembers” and tracks the trails of foraging by releasing pheromones. Once it finds food on one of the trails, other ants will aggregate to this path, increasing the amount of pheromones on the path. Ants are also able to distinguish the different distances of the paths, and will choose the nearest; at the same time, they will notify another ants by means of the amount of pheromone on this path. This phenomenon reflects a kind of swarm intelligence and collective wisdom(see Figure 8-1).

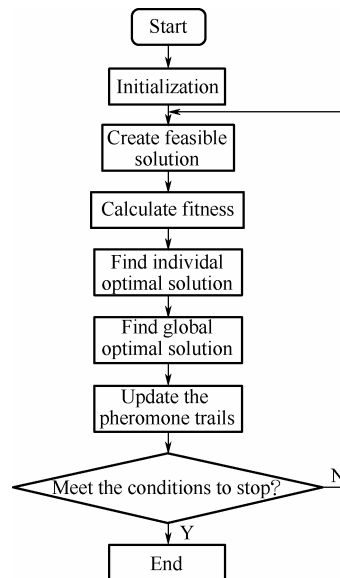


Figure 8-1 Problem solving flow chart based on Ant Colony Algorithm

In 1959, Pierre-Paul Grassé initially proposed Stigmergy theory to explain the behaviors of termites building their nests, as identified in research on Ant Colony Algorithm. In 2000, Marco Dorigo et al. published a reference review on Ant Colony Algorithm in *Science* ^[32, 33], which pushed this research field into the international academic frontier. An Ant Colony Algorithm (ACA) is a probabilistic algorithm that finds the optimal path on a map, as shown in Figure 8-1. The basic algorithm is as follows.

Algorithm 8-1
Step 1: Set the parameters and initialize pheromone trails.
Step 2: Create m feasible solutions.
Step 3: Calculate the fitness of each ant.
Step 4: Find the best position of every ant (the optimal solution).
Step 5: Find the best position on a global level (the optimal solution).
Step 6: Update the pheromone trails.
Step 7: Decide whether the termination conditions are met; if so, go to the end.
Iterate or go back to Step 3.

For example, in the TSP (Travelling Salesman Problem) which is classically solved by means of Ant Colony Algorithm, in order to simulate the behaviors of real ants, the initial parameters are set as follows^①: m represents the number of ants in the colony; n represents the number of cities; d_{ij} represents the distance between city i and city j ; $r_{ij}(t)$ represents the intensity of the pheromone trails along edge (i, j) at time t ; η_{ij} represents the visibility of edge (i, j) ; in the Ant Colony Algorithm, η_{ij} usually equals the reciprocal of the distance between city i and city j (i.e., $\eta_{ij} = 1 / d_{ij}$); $\Delta \tau_{ij}^k$ denotes the amount of pheromone per unit of the trail length that ant k left on edge (i, j) and; p_{ij}^k represents the probability of ant k transferring to city j , the city that the ant has not yet visited^[34].

Assume that the amount of information on each path is the same, and set $r_{ij}(0)=C$ (C is a constant); each ant independently chooses the next city on the path according to the retained information. At time t , the probability of ant k transferring from city i to city j , namely, p_{ij}^k , is as follows:

$$p_{ij}^k = \begin{cases} \frac{\tau_{ij}^\alpha \cdot \eta_{ij}^\beta}{\sum_{s \in \text{allowed}_k} \tau_{is}^\alpha \cdot \eta_{is}^\beta}, & j \in \text{allowed}_k \\ 0, & \text{otherwise} \end{cases} \quad (8-1)$$

α indicates the significance of the amount of information that ants accumulate in movement; β indicates the significance of the heuristic information that influences an ant's choice of the path in movement. Further, $\text{allowed}_k = \{0, 1, 2, \dots, n-1\}$ - tabu_k represents all the cities that ant k is allowed to select in the next step, where tabu_k denotes the cities to which ant k has currently traveled. When all n cities have been included in tabu_k , ant k will complete one

① For more information, please refer to the following paper: Yanling Wang, Longshu Li, Zhe Hu, Collective Intelligence Optimization Algorithm [J], *Computer Technology and Development*, 2008, 18 (8): 114-117.

cycle, and the path of ant k is then a solution to the problem. When all ants have completed one cycle, the information in each path will be adjusted according to equations (8-2) and (8-3):

$$\tau_{ij}(t+n) = (1-\rho) \cdot \tau_{ij}(t) + \Delta \tau_{ij}, \quad \rho \in (0,1) \quad (8-2)$$

$$\Delta \tau_{ij}(t) = \sum_{k=1}^m \Delta \tau_{ij}^k(t) \quad (8-3)$$

ρ represents the evaporation coefficient of the information on the path; $1-\rho$ denotes the retention coefficient of the information; $\Delta \tau_{ij}$ represents increment of the information on path ij in this cycle. If ant k has never been to path ij , then the value of $\Delta \tau_{ij}$ will be zero; otherwise, it will be Q/L_k (where Q is a constant and L_k denotes the total length of the path that ant k has travelled in this cycle).

The Ant Colony Algorithm based on the Traveling Salesman Problem proposed by Marco Dorigo in 1997 is also a good example of applying the algorithm to collective intelligence. In the Traveling Salesman Problem, a salesman must find the shortest possible route to visit many cities once given a group of cities and the distances among them. The Ant Colony Algorithm for solving the traveling salesman problem uses a virtual “ant” that explores different routes and leaves virtual “pheromones” that may gradually disappear over time, which indicates the characteristic signal that a salesman left on a path. Based on the principle that “the higher the amount of pheromone is, the shorter the route”, the best route can be determined. By simulating an ant colony’s searching for food, we effectively solve the problem of increased complexity of traveling salesmen when the number of cities increases.

3. Model of Collective Intelligence Based on Particle Swarm Optimization

Particle Swarm Optimization (PSO) is an evolutionary computation technique developed by James F. Kennedy ^[35] and Russell C. Eberhart in 1995, which was derived from the simulation of a simplified social model. The technique was originally intended to graphically simulate the beautiful and unpredictable movements of birds, and it was derived from the artificial life and evolutionary computation theories. By observing the social behavior of animals, scholars found that information sharing in a community can lead to the acquisition of greater benefits during evolution. PSO has been applied in numerous applications, including neural network training, disease medical analysis and robotic applications.

Wei Yang et al. ^[36] (2004) stated in a review of PSO that when practical problems are solved using PSO, the solution to the problem corresponds to the position of a bird in the

search space, where the birds are called “particles” or “subjects”. Each particle has its own position and speed (depending on the direction and distance of its flight), and an adaptation value is determined by the optimal function. Every particle memorizes and follows the current optimal particle, searching in the solution space. Further, each iteration of the process is not completely random: if a better solution can be found, it will become the basis for exploring the next solution.

Given that PSO involves a group of random particles (requiring a stochastic solution), the particles update themselves by tracking two “extreme values”: ① the best solution that can be found by the particle itself, called the individual extreme point (where P represents its position); and ② the best solution that can be found so far in the entire community or in a portion of the group. In the global version of PSO, the entire community is considered, and this point is called the global extreme point (where G represents its location). By contrast, in the local version of PSO, a portion of the group rather than the entire community is considered neighbors of the particles; thus, the best solution among all the neighbors is called the local extreme point (where L represents its position). After finding the two best solutions, the particles update their velocities and positions according to the relevant formula. The information of particle i can be expressed with a D-dimensional vector, where the position is $X_i = (x_{i1}, x_{i2}, \dots, x_{iD})^T$, and the velocity is $V_i = (v_{i1}, v_{i2}, \dots, v_{iD})^T$; other vectors are similar.

The basic procedures in PSO can be described as follows:

(1) Initialize the particles: position X_i^0 and speed V_i^0 of the initial search point are usually generated randomly within the allowable scope; the P coordinates of each particle are set to its current position, and the corresponding individual extreme value (the fitness value of the individual extreme point) is calculated. The global extreme point (the fitness value of the global extreme point) is the optimal individual extreme point; the serial particle number of the optimal value is recorded, and G is set as the current position of the optimal particle. As an example, Russell C. Eberhart and Xiaohui Hu (1999) proposed an analytical model for tremor behavior based on PSO following relevant theories and previous findings^[37], and in the phase of initialization, they set 30 particles and their current positions; then, they set the maximum tremor rate (as the initial search point and its speed).

(2) Evaluate the particles: this step mainly involves calculating the fitness values of the particles in order to update the individual extreme values and global extreme values. Russell C. Eberhart and Xiaohui Hu (1999) referenced previous studies on tremor

behaviors to evaluate each particle's behavior and calculate the fitness values of the particles. If the value for a particle is superior to the current individual extreme value, then this value is set as the position of this particle, and the individual extreme value is updated. If the best value among all the particles' individual extreme values is superior to the current global extreme value, then this value is set as the position of the particle; the serial number of the particle is recorded, and the global extreme value is updated.

(3) Update the particles: update the speed and position of each particle with the updating equation. According to the previous evaluation of particles' tremor behavior, certain rules are adopted to update the particles. For example, in the study on tremor behaviors based on PSO, Russell C. Eberhart and Xiaohui Hu (1999) updated the particles by regulating the weight of each layer of the neural network based on the tremor behavior.

(4) Verify whether the termination conditions are met: if the number of current iterations reaches the preset maximum number (or if the minimum number of false requests is met), then the iteration is stopped, and the optimal solution is identified; otherwise, step 2 is repeated. For example, in this case, the initial damping weight of 0.9 decreases to 0.4 after 2000 iterations; thus, the iteration is stopped, and the optimal solution is output. Russell C. Eberhart and Xiaohui Hu (1999) found that the neural network model based on the PSO algorithm can effectively distinguish between normal people and tremor patients. In the application above, this method can be used to make quick, accurate diagnoses. Further, in the diagnosis of diseases such as breast cancer and heart disease, a PSO-trained neural network can also achieved a higher diagnostic success rate, demonstrating the merits of PSO in generating collective intelligence.

8.2.6 Simulation of Collective Intelligence in Social Networks^{①②}

1. Wikipedia

Wikipedia, in which entries are created by Internet users through their collective participation, is a typical case where collective wisdom is generated through the participation, editing, interaction of the public; therefore, it provides high-quality data

① This section is based on the following papers: Shiyu Du, Jiayin Qi. Modeling and Simulation on Collective Intelligence in Future Internet - A Study of Wikipedia [J]. Information Technology Journal, 2013, 12 (20): 5531-5535. Bo Huang, Shiyu Du, Jiayin Qi, Research of Collective Intelligence in the Future Internet Based on Wikipedia Entry Classification [J], Journal of Computer Applications, 2013, 30: 48-49.

② Reference link: <http://pan.baidu.com/s/1bnrgs0J>.

resources for the study of collective intelligence. Collective intelligence advocates such concepts as basic openness, equivalence, sharing and global operation. First, in terms of openness, there are 282 different language editions of Wikipedia; as of January 2013, the total number of registered users exceeded 32 million, and the total number of edits exceeded 1.2 billion; both users and entries on Wikipedia are on the increase; the users are diverse, independent, decentralized, scattered, and widely distributed. Secondly, in terms of equivalence, Wikipedia abandoned the hierarchical model, to encourage users' self-organized production and development. In general, Wikipedia users are given the autonomy to edit any entries, while professional editors are responsible to eliminate maliciously damaged entries and accounts. These measures serve to improve user's preferences and increase the quality of entries. Thirdly, in terms of sharing, Wikipedia's sharing feature guarantees the increase of participants and the generation of UGC, while improving user's preferences for generating and sharing UGC. Finally, in terms of global operation, with people from all parts of the world participating in the collaboration, such a broad distribution range also constitutes an advantageous condition to ensure the enormous number of participants.

To constitute data for this study^①, on the one hand, we selected the time-series data of the numbers of both users and entries, the number of edits, users' demographic data and other related data spanning from January 2001 to December 2011 provided on the open platform of Wikipedia; on the other hand, on the Chinese Wikipedia, we selected the new item "entry score" as the indicator of UGC quality, which is composed of four parts: credibility, objectivity, completeness and readability. The total score of each part is 5 points, and the sum of the four parts (20 points) is the final score of an entry. Table 8-2 shows an example of the obtained data. In addition, entries on Wikipedia are divided into 11 categories: culture and art, people, geography, social and social sciences, history and events, natural and physical sciences, technology and applied science, religion and belief, health, mathematics and logic and philosophy. Due to the huge number of sample entries, it is important to select an appropriate sampling method. In terms of overall distribution, the number of edits of the sampled entries conforms to the Power Law distribution. If an ordinary, random sampling method is used, it will probably result in a large number of entries with the number of edits of 1 or even 0; these results with respect to the number of edits are

① Reference link: <http://stats.wikimedia.org/ZH/ChartsWikipediaZZ.htm>

not significantly different from each other, which cannot reflect the impact of the number of edits on the quality of entries. Therefore, the stratified sampling method is adopted: first, samples are stratified according to the number of edits, followed by sampling pro rata from each level. It is predictable that the higher the number of edits that the entries in a level have, the higher sampling ratio the level is given. Thus, entries are distributed into different hierarchies based on the number of edits, and then simple random sampling is conducted from these entries in each hierarchy, to achieve entry samples with a capacity of 105. Table 8-2 shows some examples of the 105 samples.

Table 8-2 Example of sample data for entry scoring

Number of edits	Entry Name	Entry Score
1698	Hong Kong Disneyland	15.3
1684	Zhejiang Province	16.2
1653	Detective Conan	15.5
409	Photorespiration	15.5

2. Modeling Collective Intelligence in Social Networks

Based on the above discussions, the hypotheses and assumptions presented in this model are as follows:

Hypothesis 1: as a present entry is edited and modified by Internet users based on the existing knowledge and quality of the entry, the more frequent an entry is edited or modified, the higher the quality of it.

Hypothesis 2: the diversity of users is directly related to the diversity of UGC which they produce; the more diversified the users are, the more diversified the entries in the network.

Assumption: provided that the above hypotheses are well supported, the larger the user size, the higher the overall quality, the total number, and the variety of UGC; the higher the level of collective intelligence, the more users will be attracted to participate in group collaboration.

Based on the multi-agent modeling method, the specific composition of the collective intelligence model based on Wikipedia in this study is as follows.

The agent layer: because Wikipedia users participate in entries generation and editing in a self-organizing manner and on an equal footing, we assume that each agent has the

same attributes, and multiple homologous agents are abstracted into one agent class, which is defined as *InternetUser*.

The individual agent attribute model layer: this layer is composed of four parts: inner states, sensors, effects, and environment.

- Internal states: User-Attributes=(intXPos,yPos,double IUknowledgeLevel,intIUvariety), where “intXPos” and “yPos” represent user’s location in the network; “double IUknowledgeLevel” represents the educational level of an agent; “intIUvariety” denotes user category. As 11 categories of entries are defined in Wikipedia, 11 different UGC classes are also defined correspondingly.
- Sensors: defined as “viewUGC()” in this model.
- Effects: defined as “generateUGC()” and “editUGC()” in this model.
- Environment: comprises three output variables in this model, to show the representation of collective intelligence: UGCquality (the quality of entries), generateUGCcacc (the quantity of entries), and calculateVar (the variety of entries).

The MAS layer: according to real data on Wikipedia, three behavior rules and four data rules are defined as follows:

Behavior Rule 1: a viewed entry, if user interaction (agent overlapping) or if the entry quality is greater than 16.

```
if ((IUSpace.getObjectAtX$Y (xPos,yPos)!=null)||internetEnv.UGCquality>16){ viewUGC
();}
```

Behavior Rule 2: an editable entry, if the entry quality is less than 20.

```
public void viewUGC ( ) {generateUGC( ); if (internetEnv.UGCquality<20)
{editUGC();}}
```

Behavior Rule 3: an agent can choose to move in one of the eight directions around it at the rate of once every second; overlapping may occur in the process of movement.

```
xPos += Globals.env.uniformIntRand.getIntegerWithMin$withMax(-1, 1);
yPos += Globals.env.uniformIntRand.getIntegerWithMin$withMax(-1, 1);
xPos = (xPos + modelswarm.worldX) % modelswarm.worldX;
yPos = (yPos + modelswarm.worldY) % modelswarm.worldY;
```

Data Rule 1 is shown in Table 8-3.

Table 8-3 Fitting functions for changes in agent number

Time Period	Fitting Equation	Parameters	Goodness of Fitting
2001.1~2005.11	number of agents = $\alpha \times \exp(\beta \times t)$	$\alpha=0.3023, \beta=0.2016$	0.9981
2006.1~2007.1	number of agents = $at + \beta$	$\alpha=39.82, \beta=-1082$	0.9974
2007.3~2011.11	number of agents = $(\alpha - 0.3 \cdot t) \cdot t + \beta$	$\alpha=67.94, \beta=-1708$	0.9996

Data Rule 2 is shown in Table 8-4.

Table 8-4 Fitting functions for numbers of new UGC

Time Period	Fitting Equation	Parameters	Goodness of Fitting
2001.1~2007.1	number of new entries = $\alpha \times (\text{no. of par.}^\wedge \beta) + \gamma$	$\alpha=26990, \beta=0.417, \gamma=16330$	0.9493
2007.3~2011.11	no specific fitting function in this period, during which data fluctuated between 7000~10000		

Data Rule 3 is shown in Table 8-5.

Table 8-5 Fitting functions for number of edits

Time Period	Fitting Equation	Parameters	Goodness of Fitting
2001.1~2005.1	number of edits = $\alpha \times t^\wedge \beta + \gamma$	$\alpha=2.118, \beta=0.4201, \gamma=2.889$	0.9288
2005.3~2011.11	number of edits = $\alpha \times t^\wedge \beta + \gamma$	$\alpha=1.139, \beta=0.4653, \gamma=7.39$	0.9993

Data Rule 4 is shown in Table 8-6.

Table 8-6 UGC quality fitting functions

Fitting Equation	Parameters	Goodness of Fitting
UGC quality = $(\alpha \times x^\wedge 3 + \beta \times x^\wedge 2 + \gamma \times x + \lambda) / (x + \mu)$	$\alpha=-4.96\text{e-}008, \beta=0.0008735, \gamma=14.02, \lambda=245.9,$ $\mu=29.8$	95% confidence interval

3. Model Simulation and Results Demonstration

On the Swarm platform, with the above fitting functions being the behavior rules of agents, the relationship between the size of a group and the quantity, quality and variety of entries is achieved through the simulation results of the model.

As the changes in Internet users are divided into three stages, three simulation points (25s, 35s, 45s) are selected respectively from the three stages to reflect changes in the results. As shown in Figure 8-2, Figure 8-3, and Figure 8-4, when the simulation time is 25s, the number of agents is 47000, the cumulative number of UGC is 90510, and the overall UGC quality is 14.89. When the simulation time is 35s, the number of agents is 312000, the total number of generated UGC is 287089, and the overall UGC quality is 15.66. When the simulation time is 45s, the number of agents is 742000, the total number of generated UGC is 2122030, and the overall UGC quality is 16.20. At the same time, as shown in Figure 8-5, we can tell from the outputs of Java Eclipse command lines that when the simulation time is 15s, the entry varieties 2, 3, 7, 8, 10 and 11 (Wikipedia has 11 entry categories) do not exist; when the simulation time is 25s, the entry varieties 10 and 11 do not exist; when the simulation time is 35s, all the entry varieties exist. The quantity, quality and variety of visible entries all increase with the increase of the group size.

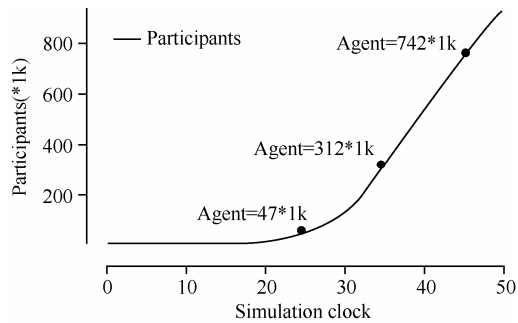


Figure 8-2 Agent quantity change trend chart

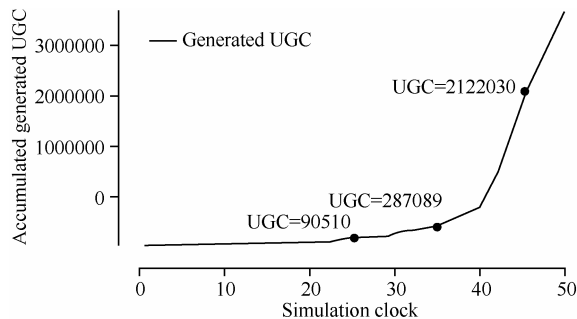


Figure 8-3 UGC quantity change trend chart

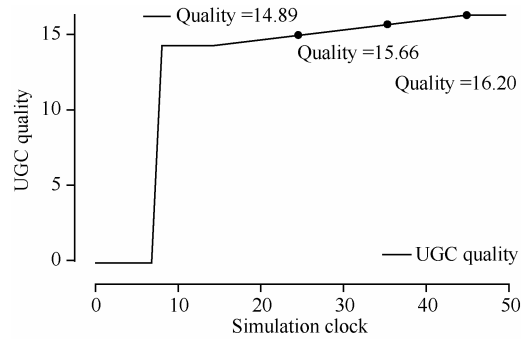


Figure 8-4 UGC quality change trend chart

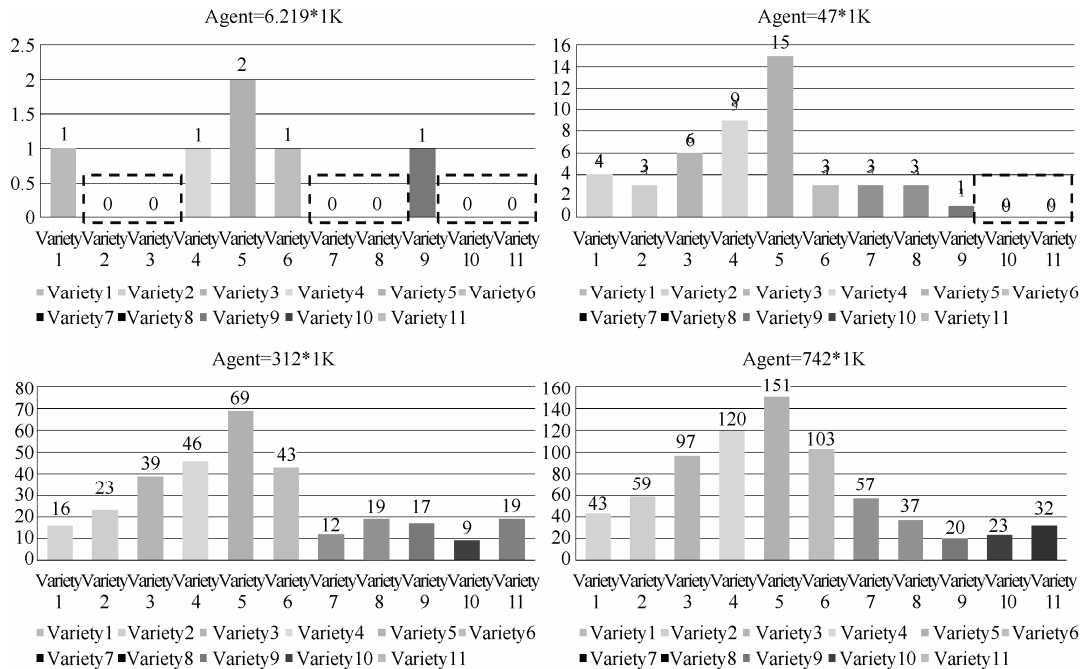


Figure 8-5 Entry variety in groups of different scales

4. Collective Intelligence Simulation Taking Into Account Entry Classification

In order to further explain how group participation forms a relationship with entries editing and professional knowledge, it is necessary to explore whether or not collective intelligence is equivalent to knowledge. On the basis of the above research results, this study intends to classify the 105 sample entries by “knowledge” or “information”.

According to the knowledge oriented view of information, information is the raw material of knowledge, while knowledge is the abstracted products of information processing. As shown in Table 8-7, the entries “Hong Kong Disneyland” and “Detective Conan” are classified into the “information” category, because such entries have relatively low threshold of editing, which are inclined to information transmission; the entries “Zhejiang Province” and “photorespiration” are classified as “knowledge” entries, because these entries have significantly higher threshold of editing, requiring professional knowledge on geography, history, biology and other disciplines, which are refined information products. 59 out of the 105 sample entries fall into the knowledge category, and 46 into the information category.

Table 8-7 Reclassifying entries into information and knowledge

Number of Edits	Entry Name	Total Score	Category
1698	Hong Kong Disneyland	15.3	Information
1684	Zhejiang Province	16.2	Knowledge
1653	Detective Conan	15.5	Information
409	Photorespiration	15.5	Knowledge

Take the above two types of entries respectively as the fitting of the number of edits and the score of entries, to achieve the relevant fitting equation as shown in Table 8-8.

Visualize the above mentioned equation with MATLAB or other mapping tools, to generate graphs which show the relationships between the quality of entries, either in the knowledge or the information category, with the number of edits, as shown in Figure 8-6.

Table 8-8 Fitting results in the condition of entry sample classification

	Fitting function: $f(x) = (p_1 \cdot x^3 + p_2 \cdot x^2 + p_3 \cdot x + p_4) / (x + q_1)$
	Parameter estimation (95% confidence interval)
	$p_1 = -1.268\text{e}-008 (-3.398\text{e}-007, 3.145\text{e}-007)$
	$p_2 = 0.0002874 (-0.001616, 0.002191)$
	$p_3 = 14.8 (12.03, 17.57), p_4 = 615.9 (-2468, 3700)$
	$q_1 = 65.62 (-209.3, 340.6)$

(To be continued)

Continued table

Information	Fit Goodness Index SSE: 92.33
	R-square: 0.3606, Adjusted R-square: 0.3132, RMSE: 1.308
	Fitting function: $f(x) = p_1 \cdot x^3 + p_2 \cdot x^2 + p_3 \cdot x + p_4$
	Parameter estimation (95% confidence interval)
	$p_1 = -6.737e-011 (-4.453e-010, 3.106e-010)$
	$p_2 = 5.803e-007 (-1.637e-006, 2.798e-006)$
	$p_3 = -0.000142 (-0.003909, 0.003625)$
	$p_4 = 13.94 (12.06, 15.83)$
	Fit Goodness Index SSE: 73.9
	R-square: 0.47, Adjusted R-square: 0.4322, RMSE: 1.326

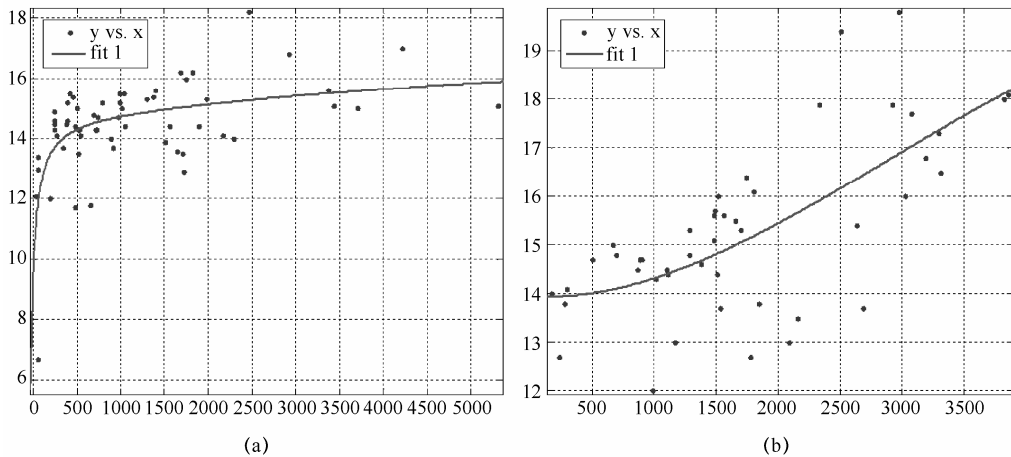


Figure 8-6 Relationship between quality of entries and number of edits (knowledge entries/information entries)

It can be seen from Figure 8-6 that the quality of both knowledge entries and information entries improves with the increase of the number of edits. The difference between the two types lies in the fact that, as shown in Figure 8-6 (a), the quality of knowledge entries improves slowly and the goodness of fit is relatively low ($R=0.3606$). In contrast, as shown in Figure 8-6 (b), the quality of information entries improves fast and the goodness of fit is relatively high ($R=0.47$). Therefore, such a difference shows that the

result achieved from information entries is more convincing than knowledge entries in supporting the conclusion that “the increase of the number of edits and that of the group size will lead to the increase of the quality of entries”. In addition, the increase of group size exerts greater impact on the quality improvement of information entries than on that of knowledge entries.

Therefore, we can explain the relationship between group participation and the formation of professional knowledge in this way: the role of group participation in the formation of collective intelligence is more significantly reflected in the information level, while the formation of professional knowledge might require the intervention of more expert wisdom.

5. Result Analysis

This model studies the influence of group size on collective intelligence based on the generation of collective intelligence. Through the collection and research of actual data on Wikipedia, we came to the following findings:

(1) The size of the group plays a decisive role in forming collective intelligence in the Internet. In this study, we found that the larger the group size is, the higher the total number, the overall quality, and the variety of UGC. The higher the level of collective intelligence, the more users will be attracted to participate in group collaboration. When the population size reaches 400000, the number of UGC begins to increase exponentially, UGC quality begins to exceed 16, UGC variety reaches the maximum value, and group collaboration begins to generate collective intelligence. In addition, the quality of information entries also increases with the increase of group size. Finally, the role of group participation in the formation of collective intelligence is more significantly reflected in the information level, while the formation of professional knowledge might require the intervention of more expert wisdom.

(2) In particular, this study explores the nature of collective intelligence and finds that information entries are more likely to be positively influenced by the growth of group size than knowledge entries, which indicates that the collective intelligence formed on the basis of group participation inclines to the transmission of information rather than the production of knowledge; the formation of knowledge may require the participation of experts. This conclusion may remind people of the research results concerning the comparison between collective intelligence and individual intelligence published by Woolley et al. in “Science” in 2010, which shows that as far as information-based applications like video games are

concerned, where the ultimate victory is achieved by converging group provided experience, such as, game strategies, paths and equipment; in these cases, collective intelligence is significantly superior to individual intelligence. In contrast, in the case of knowledge-based disciplines such as architectural design, collective intelligence is not so significant any more, where the ultimate goal is attained mainly through individual's expertise. Naturally, this is because the information needed for playing video games is evidently less professional than the knowledge or information needed for architectural design.

This study attempts to make an exploratory answer to how humans can increase their collective wisdom; that is, human beings can improve their intelligence through large-scale interactions; however, intelligence achieved in this way is more significantly embodied in the information level; as for intelligence achieved in the knowledge level, expert intervention is required. In addition, after the group size exceeds a certain value, the growth rate of entry quality turns modest, indicating that even though collective intelligence still exists, a group size of billions class is required in order to push the entry quality to the level of 20. Such a situation is, on the one hand, due to the fact that the sampled data is rated by users, which is more or less subjective, as users barely give an all-round full-score evaluation on an entry; on the other hand, it also shows the limitations of group intelligence itself; that is, it is difficult for group intelligence to approach the "truth" in the future due to the lack of expert participation.

8.3 Mechanisms Engendering Group Polarization

8.3.1 Group Polarization

Group polarization was first proposed by James Arthur Finch Stoner at the Massachusetts Institute of Technology in 1961^[38]. Through an empirical study, he found that in a scenario of group decision-making, individual opinions or decisions tend to produce uniform results of the group owing to the influence of discussions among group members. As a sociological concept, group polarization has a certain connection with the "herd behavior" (the bandwagon effect) in psychological and financial markets researches, and with the "information cascades" in informatics.

Herd behavior in psychology, according to the definition of Scharfstein in 1990^[39], refers to the phenomenon in which investors go against the Bayesian posterior distribution

of rational rules by merely following others in their actions, while ignoring their own private information. Being widely applied in financial analysis, the herd effect describes a kind of financial investment behavior in which investors ignore private information and choose to follow the crowd. This effect highlights that individuals are very likely to lose themselves in a group, as they tend to trust the information that is generally held by other members in the group and lose the ability to judge the value of the information.

According to research results in information science, an information cascade occurs when people observe others' actions and then engage themselves in the same actions, without any marking of their own personal information. In many cases, people are subject to the influences others in such aspects as views, shopping, political stance, activities, or technologies that people use (Sushil Bikhchandani et al., 1992) ^[40]. Information cascades show the influence of information on individuals. When individuals observe the behavior of other individuals in the group, they tend to hide their private information and exhibit the "herd behavior".

As a sociological term, an information cascade is often defined as a social phenomenon where decisions within a group are likely to be more extreme than individual decisions made separately because of the group influence. Researches in psychology and information science have examined the phenomenon of population polarization by focusing on individuals processing external information, whereas researches in sociology focused on the influenced of the group on individuals—that is, in sociological researches, group polarization is taken as a group decision-making behavior rather than individual decision-making behavior.

Gensheng Wang (2012) proposed a scientific classification of group polarization phenomena by dividing them into 4 categories.

(1) Unipolar aggregation, in which the views become completely uniform in the evolution process; namely, information cascade appears.

(2) Bi-polar fragmentation, in which users form two completely opposite views in the process of interaction, where the power of the two poles may be asymmetric but more stable.

(3) Multipolar fragmentation, in which users form many different but stable views in the process of interaction.

(4) Zero-polar dilution, in which a view or attitude was suddenly lost for some reason in the process of interaction.

With respect to the four categories of group polarization phenomena, the unipolar aggregation is usually associated with an unfavorable trend, in which a public crisis event

tends to arise. The bi-polar fragmentation satisfies the trend of an event where discussion leads to a positive direction, which pushes the group polarization towards the levels of multipolar fragmentation and zero-polar dilution. Because unipolar aggregation and the bi-polar fragmentation are more common, whereas multipolar fragmentation and zero-polar dilution represent special cases in the evolution process, this book focuses on unipolar aggregation and bi-polar fragmentation.

In summary, this book defines group polarization as a phenomenon in which individuals imitate others' behavior instead of behaving based on their own information under certain conditions, which causes individuals in a group to hold the same view; such behaviors can be marked by either unipolar aggregation or bi-polar fragmentation.

In recent years, the Internet and online social media have gradually become a context for observing the phenomena of group polarization. When a group of individuals begin to hold the same view of a topic and produce similar dialogues, group polarization can be observed. This book defines group polarization in a social network as a phenomenon in which individuals imitate others' behavior instead of behaving based on their own information within the context of a social network under certain conditions, which causes individuals in group to hold the same view; such behaviors can be marked by either unipolar aggregation or bi-polar fragmentation.

8.3.2 Social Comparison Theory and Group Polarization

Many theories have been proposed to explain group polarization. Among them, the most classic and comprehensible theory is social comparison theory. In 1954, Leon Festinger, an American social psychologist, proposed social comparison theory, which suggests that in the absence of objective evaluation, each individual evaluates himself or herself through comparison with others, eventually leading to convergent behavior in the group ^[41]. Festinger suggested that social comparison theory explains the reasons why people imitate the models in media; one of the reasons is to enhance individuals' self-confidence which becomes the reasonable basis for self-perfection.

After Festinger proposed the social comparison theory, many new theories have expanded and continuously improved researches on social comparison. For example, Mackie et al. pointed out that the theory proposed by Festinger ^[42-45] that "social comparison mainly involves the comparison with those who are similar to themselves" was not comprehensive. In social comparison in interpersonal communication, on the one hand, people make similar people the

object of comparison to confirm that they are similar to others; on the other hand, they compare themselves with different people to confirm their own abilities from the opposite side in order to improve their self-evaluation and develop their own social behavior, which serves as the auxiliary social comparison. A wise man would combine the two aspects of social comparison in order to improve his or she self-evaluation. Therefore, in 1979, Tajfel Henri and John Turner proposed three stages of evolution in social comparison in which individuals evaluate both ego and alter ego. The three stages of the social comparison theory are: self-categorization, social identity and social comparison^[46].

Self-categorization refers to the process of establishing a psychological relationship between the individual and the group. Then, the individual can psychologically become a member of the group and form relationships with other members of the same genus. The central element of the cognitive process of self-categorization is depersonalization. Briefly, depersonalization refers to the phenomenon in which people deliberately maintain a distance from others who belong to different types of groups and even take an attitude of indifference and hostility. Social identity is the process through which an individual is judged to be a member of a particular group through subjective perception, forming the in-group and the out-group. The final stage is social comparison, where small groups begin to form owing to the increased homogeneity in the two groups. That is, the extreme views of the groups become intensified, leading to the formation of group polarization. Hence, social comparison theory explains the mechanism through which group polarization takes shape. The theoretical formation process is shown in Figure 8-7.

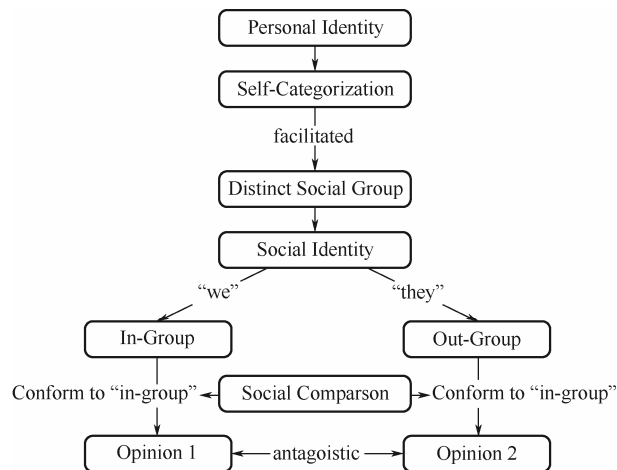


Figure 8-7 The evolution of social comparison theory

8.3.3 Conditions Engendering Group Polarization

Tom Hayes described the following cases in his book *No Size Fits All*. In December 2007, a research on MovieLens, an online film review community, shows that when a user knows about the behavioral mode of other people, he tends to intentionally behave in conformity with the norm of the group, resulting in identified behavior. In the research, the researchers merely sent the users a message about the average usage level of users on the site. After the sending of the message, the monthly comments from low active users increased by 530%, while the monthly comments from high active users decreased by 62%. That is to say, all the users changed their behaviors to make it closer to the median value of the group.

This case highlights the view of Burns, one of the researchers who developed cognitive therapy: when a group exerts some pressure on its members, the advantages of group thinking will not exist. This phenomenon is also known as the information cascade effect. The core value of the effect is that people tend to naturally abandon their capacity of independent thinking and follow the practice of others, when the collectively identified authenticity exceeds a certain critical value.

An information cascade forms in the following four stages: 1. A triggering event emerges. 2. An idea is proposed by someone who is believed to be trustworthy. 3. The rest people notice this idea and believe that agreeing with this idea and giving up their own independent thought is in their best interests. 4. The more people think that this idea is correct, the more people and new information receivers will also think that it is true.

In the book *Science and Practice*, Robert B. Cialdini, the world's leading expert of influence research, also noted that the more people think a certain point of view is correct, the more correct the viewpoint is ^[47]. Moreover, he proposed that social certification is particularly influential in two cases. First, when people face uncertainty (i.e., when they do not have sufficient information and do not know the truth), they tend to observe other people's practices and views and think that others' behavior is correct. Secondly, in order to be identified with others, people tend to follow and imitate the practices of people with whom they have higher homogeneity. Furthermore, Owen Janis put forward five important characteristics of a homogeneous group: limitation of discussions among group members, the superiority of majority-supported views, the neglecting of minority's views, the repellency against external views, and the convergence of group views. These results

trigger further exploration into the mechanisms which give rise to group polarization.

Domestic scholars have also conducted relevant research on the conditions that engender group polarization. For example, Bo Shi (2010) noted that there are three necessary conditions for the generation of group polarization: the emergence of an event, network filtering and group collaboration^[48]. In contrast to real society, a network provides a natural environment for systematic information filtering, which makes it easier for network groups to develop internal homogeneity and intra-group heterogeneity. Further, group cooperation is determined by the nature and characteristics of the network group. Individuals in the group, through the mechanism of group consciousness, experience an essential subconscious psychological change and thus lose their self-consciousness and converge toward the view of the group.

Based on the above description, we can conclude that there are four conditions that engender group polarization: First, there must be a triggering event. Secondly, individuals within the group know about the choices of others. Thirdly, there is a lack of group information. Fourthly, there is a certain level of homogeneity in the group; through collaborative filtering, people choose to follow the views of others without thinking independently and believe that this is the best choice.

8.3.4 Factors That Influence the Formation of Group Polarization

The factors that influence the formation of group polarization can be viewed from three dimensions: subject, group and information.

1. Subject Dimension

In discussing “group classification”, Bon suggested that when individuals are members of a group, their group psychology differs from their individual psychology, and their ability of individual thinking is influenced by such difference. Thus, when an individual is no longer independent, the phenomenon of group polarization is very likely to arise. Social identity theory holds that when individuals identify themselves with a group, they will consciously conform their own attitude to the group recognized by them. Further, American political scientist Zaller examined how individuals in a group are exposed to information, how they receive information, and how they form a public opinion. Domestic scholar Xinzhou Xie (2004) stated that if an individual’s ability of independent thinking is strong enough, he or she will be more tolerant to other independent-minded views instead

of blindly following them ^[49]. Thus, individuals' degree of independence has a very important influence on group polarization.

Further, Cass R. Sunstein (2002) believes that, according to the "first possession theory" and "the first impression theory", all individuals subconsciously have their own self standards; first possessed opinions or first impressions often play a dominant role in their minds, and it is not easy for people to change their viewpoints in front of different opinions ^[50]. Thus, individuals' initial views have influence on the change of their subsequent views.

In summary, we can draw a conclusion from the subject dimension that the following factors affect population polarization: the degree of independence and initial opinions of individuals.

2. Group Dimension

Le Bon Gustave, a French scholar, stated that individual's intelligence and his personality are both weakened in collective psychology. Heterogeneity is swallowed by homogeneity, and the quality of unconsciousness is in the upper hand. Further, Cialdini stated that people tend to follow and imitate people whose behaviors are similar to their own. Related researches have shown that group pressure on an individual is greater in a group where an individual is active than in a group with which the individual is completely unfamiliar. Moreover, at the group level, group polarization is more likely to arise in a group of acquaintances than in a group of strangers. Further, Morris Charles (2000) showed that when a cascade encounters a high-density cluster (if each node has at neighbors at a proportion of N that also belong to the node set, then this node set is called a cluster with a density of P), it will stop, and this is the only reason why the cascading stops ^[51]. Thus, high homogeneity within the group is required for population polarization to arise.

From the perspective of economics and finance, Harrison Hong (2005) found that fund managers are vulnerable to the influence of fund managers in both the same city and other cities; that is, the herd effect exhibits when fund managers select stocks ^[52]. Accordingly, intensive interaction among people is a prerequisite for the generation of group polarization. Further, Yang Shanlin et al. (2009) found that the duration of interactions and the size of the group had a significant impact on the emergence of herd behavior, which follows certain rules. Thus, the interaction duration and the group size are factors that influence group polarization ^[53]. We can then conclude therefrom that the interaction frequency (the density of the group) is related to the effect of group polarization.

The theory of Duncan J Watts and Sheridan Dodds Peter (2007) provokes more

thoughts^[54]. Researchers found that large-scale cascade formation is driven not by opinion leaders but by key individuals who first show polarization. Of course, opinion leaders may also be such key individuals. Thus, group polarization is triggered by both opinion leaders and the individuals in the group who first show polarization.

In summary, we can draw the conclusion from the group dimension that the following factors influence group polarization: group homogeneity, group density and the initially influenced groups.

3. Information Dimension

Haewoon Kwak and other related researchers have found that groups tend to share two types of information^[55]: the first type is the primary disseminated information, and the second type is information that requires a second thought before dissemination. The speeds with which different types of information is diffused are different. For example, for information on events (usually explosive news or headlines), the speed of diffusion within the group is very fast, as it requires little thoughts or analysis from the part of the group members. By contrast, the discussion type information (which usually carries a point of view) is diffused more slowly because it requires group discussion and review.

Group polarization is a process in which information is rapidly diffused and communicated; thus, the nature of information is very important. G. Thomas (1973) and other studies have found that sensitive topics (such as government, public policy, law, war and violent behavior) are more likely to give rise to group polarization; thus, the sensitivity of information is a factor that influences group polarization. Zhang Yiwen et al. (2012) defined two types of inner dynamism related to a sudden public crisis event: the sensitivity of the event and the public nature of the event^[56]. The sensitivity of the event is the basis for the formation of inner dynamism. The sensitivity of a topic indicates that after a crisis event occurs, it is highly possible that online public opinions will explode when certain topics are concerned. After the event, because of certain factors, the attention of Internet users, media, government and other organizations on the event will increase, which will generate a considerable amount of information on the event consequently. As information on the event gradually increases, online public opinions will form. The public nature of an event refers to the degree of its threat to the social value system. If the public nature of an event is greater, the event will pose a greater threat to the social system. When the public crisis occurs, individuals in the public will naturally consider whether such an event will affect them, and the more the public believes that similar events will affect them, the more

likely social panic will arise, and the more likely population polarization will occur. Further, Li Ke (2005) proposed that the fuzzy degree of information is an important factor in whether individuals follow the views of others. If the ambiguity of information is high, the accuracy and reliability of the information received by individuals will be very low, and the usefulness of the information will not be guaranteed. In such a case, individuals will be less likely to be convinced of a particular point of view.

In summary, we can draw the conclusion from the information dimension that the following factors influence group polarization: information sensitivity, information publicity and information ambiguity.

8.3.5 Main Models of Group Polarization Analysis

1. Herd Behavior Model Based on Game Theory and Principal-agent Theory

In real life, many activities require coordination between related people. For instance, the premise of normal bank operation is that the creditors of a bank trust that the bank is able to operate normally and reliably. Based on such a belief, they will not withdraw money from the bank at the same time. In this case, they are taking a “concerted action”. However, this trust-based “concerted action” is very fragile. Once some actors receive error messages (e.g., rumors, accidental impacts) that may make them lose faith in the action, a coordinated global collapse may occur.

In 1992, Abhijit V Banerjee proposed a herd behavior model based on game theory and principal-agent theory. The model, which is mainly used to predict economic behavior in fund and stock markets, explains this phenomenon well^[57]. In the herd behavior model, investors’ herd behavior is consistent with the rule of maximum utility and is considered irrational behavior in the presence of factors such as “group pressure”. Herd behavior models can be classified into two types: sequence and non-sequence models.

The herd behavior model proposed by Abhijit V Banerjee belongs to the sequence type; in this model, each decision maker observes a previous decision maker’s movement when making decisions. Such behavior is rational because the previous decision maker may send important signals. The decision-making process of an individual is shown in Figure 8-8.

(1) An insider is defined as a participant of a match or a game who has the decision-making power. Therefore, any participants who take part in the decision-making

process are defined as insiders in the model of Abhijit V Banerjee.

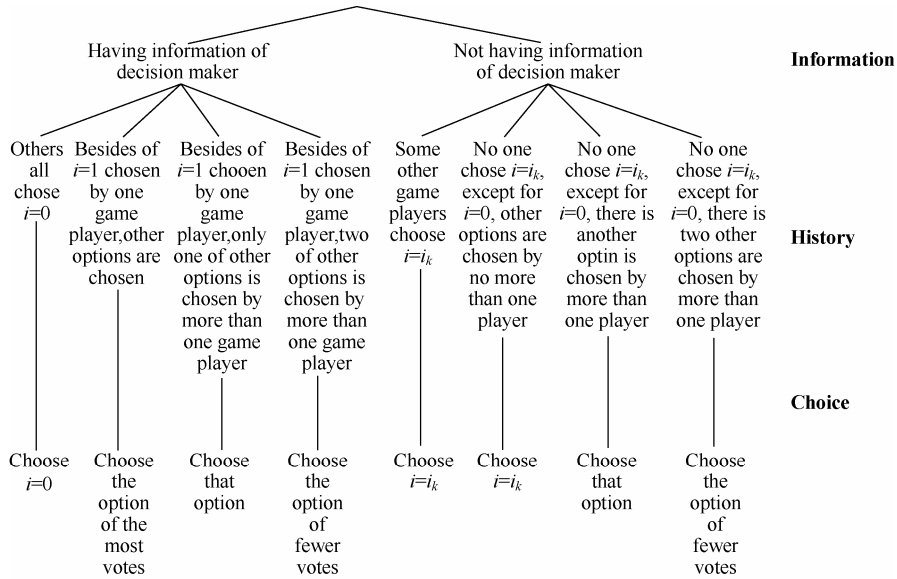


Figure 8-8 Decision-making process of decision maker k (Banerjee, 1992)

(2) Definition of game strategy: each participant in a game has to choose a practical and integrated action program which is not for a certain stage, but rather for guiding the overall action. Such a program is called a strategy of an insider. For instance, in the model of Abhijit V Banerjee, the game strategy is defined as follows: with A and B (A and B represent any two different actions that exert different effects on the actors) as any of the two elements in a set of options for this group of actors, assuming that at the beginning $N-1$ out of N actors received the signal indicating that A is superior to B, but the only one that received the opposite message which prefers B over A ranks first in the sequence of action, obviously, the first person will choose B according to the signal he or she received.

The second person, after observing the action of the first person, knows that the signal received by the first person is that B is superior to A; however, his own signal reads that A is superior to B. Under the premise that everyone is assumed to have signals of the same quality, the second person will abandon his private signal and choose B based on the prior probability action. Thus, the action of the second person does not provide new information for the next person in the sequence.

The situation for the third person is the same as that for the second person, who will make the same choice.

As the sequence proceeds, everyone will choose B instead of A even though the gathered information shows that A should be superior to B. Therefore, a herd effect arises.

If the second person always acts according to his own signal, the third person would know that the second person prefer signal A rather than signal B, and he would obviously choose A. The rest may be deduced by analogy, and the final result is that $N-1$ people except the first person will choose A over B, which exactly reflects the best choice shown by the gathered information. Because the second person abandoned his private signal to take part in the herd, a negative externality affects the rest of the people in the group. If the second person makes his decision according to his private signal, his action will provide information to the rest of people in the group. This action rule that each person in the sequence makes choices based on his own signal rather than the previous person's choice will lead the next person to act the same. In this way, the result will be different from the previous result. This type of externality is called a herd externality.

(3) Definition of gain and loss in a game: the final result of a game shows the gain and loss of a game player. Each person has his or her own result in a game—either a gain or a loss. The results depend not only on an individual's strategy but also on a set of strategies selected by all other individuals in the game. For example, in this herd behavior model, Abhijit V Banerjee assumes that a group of N actors have the same VNM^① utility functions that is risk neutral, with a target of utility maximization, when there are at least two people who have not received the right signal, the probability that the n th decision maker will receive the right signal is $1 - (1 - \alpha\beta)^{n-1} - (n-1)(1 - \alpha\beta)^{n-2}\alpha\beta$, where α is the probability of obtaining a real signal of investment profit and β is the probability that all individuals failed making the right choice. For any arbitrarily small value $\varepsilon > 0$, when $n(\varepsilon)$ is sufficiently large, the probability will be at least $1 - \varepsilon$, and the minimum value for the external utility function will be $z[N - n(\varepsilon)](1 - \varepsilon)/N$, where N is the size of the group, z is the individual's physical return on investment (see the article "A simple model of herd behavior" for more information). It's specially worth noting that when N is sufficiently large, the external utility function will approach $N(1 - \varepsilon)$. On the opposite side, if the decision maker receives the right signal, the probability of the right choice being made in herd behavior model is $IT \equiv [1 - a(1 - \beta)]^{-1}(1 - a)(1 - \beta)$, and the maximum value of the external utility function is $zN[1 - IT]$.

(4) Explain the equilibrium of the game process, and analyze the results of a game.

① VNM utility function theory is a framework for analysis established by Von Neumann and Morgenstem in 1950s on the basis of the axiomatic hypothesis by using logical and mathematical tools under uncertain conditions of rational actors.

Each game player will have a result. Equilibrium means balance, and in economics, equilibrium indicates that a related variable tends to be stable. According to the analysis of Banerjee, the characteristics of the decision made in this model are as follows: when the group is sufficiently large, the equilibrium state action is inefficient, if measured with prior probability. This explains why herd behavior is irrational. In this case, the actor has independent information, and the group is sufficiently large; thus, there always be circumstances for an individual to make the right decision. However, because the externality of herd behavior has positive feedback, the equilibrium state action is unstable in some stages of a game. The signals sent by preceding decision makers determine the formation of first group; based on this law, every individual takes part in the group action eventually.

Abhijit V Banerjee's research shows that an individual decision maker who makes decisions based on the previous actor's information seems irrational, but such behavior is rational for the individual because the previous actor may have important information that the decision maker does not know. However, following this rule may result in the Pareto inefficient equilibrium of the group behavior ^[58].

2. Group Consistency Model Based on Information Cascades

The herd behavior model proposed by Abhijit V Banerjee et al. can soundly explain the consistency of group behavior, but the spillover effect caused by different shocks and the way of sudden events causing the group behavior was not considered. In this regard, in 1992, Bikhchandani, Hirschleifer and Welch introduced the concept of an information cascade, which is one of the most important keywords in research on group behavior. The information cascade model, which is called the BHW model, was later applied in financial market analysis. The basic principle of the information cascade model is that after observing previous investors' decisions, observers decide that it is a rational and optimal choice to disregard their own private information and follow the decisions of other investors. Thus, an information cascade arises.

The following steps are used in research on behavioral consistency within groups based on information cascades.

(1) Define variables for the effectiveness of decision making according to the application scenarios and risk preference; then, refine and summarize the investment behavior of decision makers. According to the information cascade model developed by Bikhchandani et al., the investors are risk neutral. We take investments in the stock market

as an example. The investors have to decide whether or not to invest in a particular stock in turns in accordance with the order of exogenous decisions. Investment result $v \in V = \{-1, 1\}$ is randomly decided before the start of the first stage, and remains unchanged in the later stage. Good investment results are denoted as $v=1$, and bad investment results are denoted as $v=-1$. We assume that the investment results take the value $v=1$ with the probability $\mu_1 = 1/2$. The investment results are revealed after all investors have made decisions.

(2) Variational definition for the private signal obtained by decision maker, the investment decision, and the investment history observed previously based on the process of information acquisition. For instance, in the information cascade model, before making an investment decision, every investor can receive a private signal that reveals the investment results. Investors' private signals with the given investment results are independently distributed. The signal of investor i is $s_i \in S = \{-1, 1\}$. Good signals are denoted as $s=1$, while bad signals are denoted as $s=-1$. The accuracy of the private signal refers to the probability that the signal is correct with the given investment results. Except for private signals, each investor can also observe the investment decisions made by previous investors. Therefore, the information set of investor i contains his private signals and all previous investors' investment history $h_i = (a_1, a_2, \dots, a_{i-1})$. In addition, the public faith μ_i of stage i can be treated as the probability that the investment results are good, given the investment history h_i . This variable is computed as $\mu_i = p(v=1 / h_i)$. In this model, each investor chooses an investment decision that is $a_i \in A = \{0, 1\}$ in discrete sets. If $a_i = 1$, the investment decision is positive, whereas if $a_i = 0$, the decision is negative.

(3) After defining investors' returns and results, establish a model of the relationship between decision-making behavior and investment returns in combination with the process of information acquisition. Then, calculate the expected value of investors' returns according to the relevant information and signal probability. We assume that investors do not have an initial endowment. The return of investor i depends on his investment decision and investment result:

$$\mu_i(a_i, v) = \begin{cases} 0, & a_i = 0 \\ v, & a_i = 1 \end{cases} \quad (8-4)$$

In uncertain circumstances, the return of an investor is the expected value of $\mu_i(a_i, v)$ under the information set:

$$E\mu_i(a_i, v / h_i, s_i) = \begin{cases} 0, & a_i = 0 \\ E(v / h_i, s_i), & a_i = 1 \end{cases} \quad (8-5)$$

In this equation, $E(v / h_i, s_i) = p(v=1 / h_i, s_i) \times 1 + p(v=-1 / h_i, s_i) \times (-1)$. For instance, when an investor i decides to invest; namely $a_i=1$, his expected return will be $E(v / h_i, s_i) = 0.7 + 0.3 \times (-1) = 0.4$ if the probability μ_i of good signal s_i received by investor i is 0.7, whereas that of the bad signal is 0.3.

Investor i is included in this sequence decision model. The structure of the model and Bayesian rationality is public information. After observing other investors' decisions, each investor updates his investment faith in the investment results by applying the Bayesian rule and makes his investment decision.

(4) Analyze the process of obtaining information and the model results. With such an analysis, we can discuss the behavioral tendency and consistency of group decision-making based on the information cascade model and explain the causes of information cascades. The analysis of the above model shows that all the subsequent investors choose to invest when the number of the investors who choose to invest is greater than the number of investors who do not choose to invest by two or more. In such a case, the information cascade and herd behavior related to investment occurs. When the number of the investors who choose not to invest is greater than the number of investors who choose to invest by two or more, all the investors that follow choose not to invest. Thus, the information cascade and herd behavior related to giving up investment occurs. Otherwise, whether an information cascade occurs depends not only on the number of good and bad signals received by the preceding investors but also on the order of the received signals. Moreover, the information cascade is path dependent, specific and variable. In the conditions that new information is present, the accuracy of private signals improves, or the investment results change, the existing information cascade is likely to stop or change.

(5) Further analyze and discuss the polarizing effects of group decision-making and information cascades according to existing results of information cascade research. Then, using the actual problem, explain the root causes of information cascades. Bikhchandani et al. noted that an information cascade could be either positive or negative; that is, all individuals may accept or refuse a particular action. For example, a group of young people facing the decision to try drugs may have a strong motivation to try drugs if their friends are trying them; by contrast, young people may avoid drugs if their friends refuse to try them.

According to the relevant rules regarding decision making, the probabilities of

forming positive information cascade, zero information cascade, and negative information cascade are expressed as follows respectively:

$$\frac{1-(p-p^2)^{n/2}}{2}, (p-p^2)^{n/2}, \frac{1-(p-p^2)^{n/2}}{2} \quad (8-6)$$

In this formula, p represents the accuracy of private signals mentioned above, and n denotes the number of initial decision makers. When $n=2$, the probabilities of forming positive information cascade, zero information cascade, and negative information cascade can be shown as follows respectively:

$$\frac{1-p+p^2}{2}, p-p^2, \frac{1-p+p^2}{2} \quad (8-7)$$

As shown in formula (8-7), when probability p approximates $1/2$, the information cascade is delayed, and the signal provides no information. When p departs from $1/2$, the noise in the signal increases; in other words, people will more clear about whether or not to adopt a signal or an action. Furthermore, we can conclude from the above formula that the probability of zero information cascade occurring exponentially decreases as the value of n increases. For instance, for a signal, if $p=1/2+\varepsilon$ and if ε is assumed to be arbitrarily small, the probability of zero information cascade occurring is less than 0.1 even when n is set to 10.

Therefore, the probability of an information cascade occurring varies with changes in p and n , as shown in Figure 8-9.

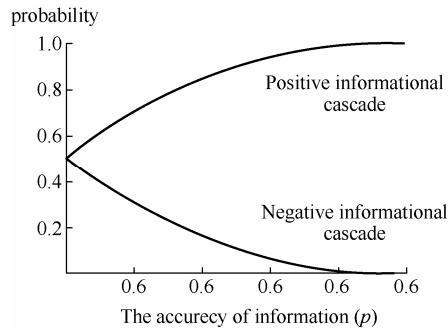


Figure 8-9 The probabilities of positive and negative information cascades occurring

The BHW model explains not only why people demonstrate uniform behaviors but also why the behavioral convergence may be idiosyncratic and fragile (i.e., eventually there will be multiple equilibria that are unstable). This model is thus one of the most important developments in research on herd behavior. Within the BHW model, we argue that the

following four kinds of mechanisms may lead to uniform behavior within a group: sanctions on deviants, positive payoff externalities, conformity preference, and communication. It explains why social behavior is of historical dependence in the former 3 mechanisms, and we can see from the last one that if communication is costless and reliable, group convergence behavior is likely to occur.

In conclusion, the important contribution of the BHW model is that it dynamically imitates information cascades in the process of decision making. Furthermore, the problem of effectively disclosing information to the public is discussed based on this model^①. People tend to have greater belief in the government when information is more accurate. Objective and adequate communication about markets and related institutions helps to prevent a coordinated collapse of the system and to influence public behavior. To perfect the information disclosure system, the government should strengthen information disclosure, increase the openness of the market in order to ensure that timely, complete and accurate information can be provided, and that the release of information can help investors to form rational expectations.

3. Simulation of Herd Behavior in Group Decision-making Based on Cellular Automata

In traditional modeling methods, complex and classical mathematical formulas always need to be established. Considering the constantly emerging complications and problems related to human behavior, it is difficult to use mathematical symbols to describe human behavior in many realistic problems. In the fields of computer science and artificial intelligence, simulation has attracted people's attention as an experimental method to explore the inherent nature of human behavior. There are currently mainly the following microscopic simulation models that obtain macroscopic results based on micro mechanisms through the simulation of individual behaviors and interactions: the multi Agent model, the seepage flow model, the critical value model and cellular automata etc.

A cellular automaton is a discrete mathematical model of time and space. This model was proposed by Von Neumann and his partners for the simulation of self-replicating biological system in the 1950s. Structurally, a cellular automaton is a Quintet, consisting of five basic parts: a cellular automaton, cellular space, neighbors, cellular automata state set and evolution rules. The main content of using applied cellular automata to build

① The integrity and sufficiency of information disclosure are necessary and sufficient conditions for the effectiveness of stock markets; further, the integrity and sufficiency of information disclosure are objective requirements for not only listed companies but also market regulation.

simulation models for actual systems is to complete the establishment and construction of the above mentioned five components by using cellular automata. The last two of the five components, the cellular automata state set and the evolution rules, are of utmost importance.

The simulation of herd behavior in group decision making based on cellular automata developed by Shan-lin Yang et al. in 2009 revealed the effect of pseudo herd behavior^① on group decision making. The simulation results show that conformity preference, the number of group interactions and group size exert significant influences on three kinds of herd behaviors, and such influences follow certain rules.

The process of simulating herd behavior in group decision-making based on cellular automata comprises the following steps.

(1) First, clearly define the concepts in the research question, and propose hypotheses of the model based on scenarios. For example, in this model, Shan-lin Yang (2009) defined the concept of herd behavior and divided group behavior into three categories: “real herd behavior”; “pseudo herd behavior” and “real non-herd behavior”; “pseudo non-herd behavior”. Based on the prerequisites of herd behaviors, two hypotheses are put forward.

Hypothesis 1: The decision making of other decision makers in the group is observable.

Hypothesis 2: All decisions are made in a sequence rather than at the same time.

(2) Based on the research hypothesis, and combined with the cellular automata model of five basic components, the cellular automata, cellular space, neighbors, cellular automata state sets and evolution rules are expounded, so as to complete the construction of the model. The two most important components among others: the cellular automata state set and the evolution rules, are defined.

For example, the evolution model and the quintet structure of cellular automata constructed by Shan-lin Yang et al. (2009) is as follows:

$$A = (d, Ld, N, S, F) \quad (8-8)$$

Where, A is the model for evolution; d represents the cellular automata; Ld represents cellular space, which is a network system composed of all individuals in the decision making group, denoted by an $n \times n$ square cellular space; each cellular automaton in the square grid indicates a decision maker; the distances between cellular automata are not

^① Many decision makers have similar information sets, therefore, they make similar decisions, which are, however, not recognizable to the market; these decision makers are mistaken for taking a herd behavior, and such a behavior is called “pseudo-herb behavior”.

geographic distances; rather, they can be taken as professional distances, or decision resources owned by other cellular automata, or the availability of the information set; N represents neighbors; S represents cellular automata state set; F represents the evolution rules. The cellular automata state set and the evolution rules are particularly expounded.

There are two types of cellular automata state sets: the mother state set S_m and the derivative state set S_c , the mother state set $S_m = S_1 \times S_2$. In this equation, S_1 denotes the conformity preference, including the “follow-the-fashion” type (δ_{11}), the environmental adaption type (δ_{12}) and the independent action type (δ_{13}). That is, $S_1 = (\delta_{11}, \delta_{12}, \delta_{13})$. Further, S_2 is the decision making resources owned by the central cellular automaton; namely, the unit amount of information. In the derivative state set $S_c = S_3 \times S_4 \times S_5$, S_3 denotes a true solution of a cellular automaton at a certain moment, S_4 denotes the final solution selected by a cellular automaton at a certain moment, S_5 denotes the real herd behavior of a cellular automaton at a certain moment, including “real herd behavior”; “pseudo herd behavior” and “real non-herd behavior”; “pseudo non-herd behavior”.

Regarding the evolution rules, usually the state at time $t+1$ is influenced by the state of the neighbor cell, the self state and control variables at moment t , as represented in the following equations:

$$S_r^{t+1} = F(S_r^t, S_{rL}^t; R) \quad (8-9)$$

$$S_{rL}^t = (S_{rL(1)}^t, \dots, S_{rL(k)}^t) \quad (8-10)$$

$$S \in S_1 \times S_2 \times S_3 \times S_4 \times S_5, t=0,1,2,\dots \quad (8-11)$$

Further, the evolution rules of the four states of the central cellular automaton, S_2, S_3, S_4, S_5 , are expounded. Due to space limitation, we will not go into details about the specific analysis process^①.

(3) When the model is established, carry out simulation on a relevant simulation platform (such as, Visual Basic, MATLAB, Swarm, etc.) through program design and implementation^②. For example, Shan-lin Yang et al. (2009) used the Visual Basic 6.0 development tool in a Windows environment to implement the simulation model. The analog input parameters include the size of the grid, group conformity preference ratio, cell size, type of neighbors, alternative sets and initial decision resources. The number of conformity preferences is set to reflect three kinds of conformity bias proportional to the

① For the full contents, refer to the following paper: Shan-lin Yang, Ke-yu Zhu, Chao Fu, et al. Simulation of the Group Decision Herd Behavior based on Cellular Automata [J]. Journal of Engineering System Theory and Practice, 2009(9): 115-124.

② The related learning code is available on the website[EB/OL] <http://pan.baidu.com/s/1dD3n285>.

group. At the same time, the neighbor type for this simulation region is set as Von Neumann, and the rules of information distribution is set to determine the owning amount of initial information. For example, in this paper, the supporting weights of 200 information elements for five backup solutions are calculated in the model, to achieve the weights of five backup solutions for decision targets $\{0.2104, 0.1894, 0.2104, 0.1894, 0.2062\}$. Thus, it is concluded from the calculation result that solution 1 is the optimal solution.

(4) Change the input parameters, and observe the changes of the output results, and then simulate and analyze the herd behavior under different conditions, to get the relevant conclusions of the simulation of herd behavior evolution. Yang Shanlin et al. (2009) carried out the simulation of herd behavior through the experiments of two cellular automata. They simulated and analyzed herd behaviors under different conformity preferences and different group sizes respectively. For example, the authors made three sets of experiments by changing the times of conformity preferences of input parameters $\gamma=\{1,0,0\}$, $\gamma=\{0,0,1\}$ and $\gamma=\{0.25, 0.5, 0.25\}$. With different interaction times, they achieved the ratios of “false conglomerates” and “not uncommon” in the output parameters respectively, and then conducted descriptive statistical analysis to obtain the trend of the change in the size of the pseudo herd behavior ratio and the number of interactions required to reach the optimal decision.

(5) After the general simulation of the model, the universal conclusion of herd behavior in group decision-making is drawn, to solve the problems such as the factors influencing the formation of group polarization behaviors in reality. For example, after the change and simulation of parameters in the model, Yang Shanlin et al. (2009) obtained the following conclusions: when the herd behavior appears to be fully acting in the group, the final evolutionary result of the group decision is very sensitive to the initial state; when the herd behavior appears in part of the group participants only, what influences the interaction convergence speed of the group is more than just conformity preference; the increase of the group size also significant influences the convergence speed of the group. That is to say, the bigger the group is, the slower the decision make process is. In the evolution of decision-making in groups that exhibit fully and partially herd behaviors, evident rules can be observed between the pseudo herd behavior and the change in group size.

It is possible to carry out modeling and simulation in studies of human group behavior by means of computer simulation, and thus to discover the main causes and laws of group

polarization behavior, which is an important method that deserves attention and exploration in related studies.

8.3.6 Simulation of Group Polarization in Social Networks Without the Influence of Social Network Structure

Based on the above factors that influence group polarization, this part is aimed to construct a multi-agent based model of group polarization in social networks; in addition, with the duration of group polarization (the time period required for 80% of Internet users to reach polarization) and the percentage of group polarization (the percentage of Internet users holding different views after group polarization) as the two representation factors of group polarization performance, the following three scientific questions are investigated:

- (1) How group polarization on the Internet is affected by these factors?
- (2) Which one of the factors exerts greater impact on the behavior of group polarization?
- (3) With all the above factors taken into account, what is the threshold level at which a group polarization behavior arises?

1. Model Hypotheses

The relevant hypotheses presented in this model are as follows.

Hypothesis 1: This model is built based on a specific media system; that is, the impact of such factors as media and network structure on group polarization behaviors is temporarily ignored.

Hypothesis 2: Based on the relevant theories and mechanisms about the generation of group polarization, we assume that when the group pressure perceived by individuals exceeds a certain threshold, the individuals will follow the group opinions without seeking to process their private information, due to the fact that individuals believe that such a behavior is to their best benefits.

2. The Formation of Group Polarization

In this model, we divide the process of group polarization formation into three stages, which are the latency period, the emergence period and the formation period, as shown in Figure 8-10.

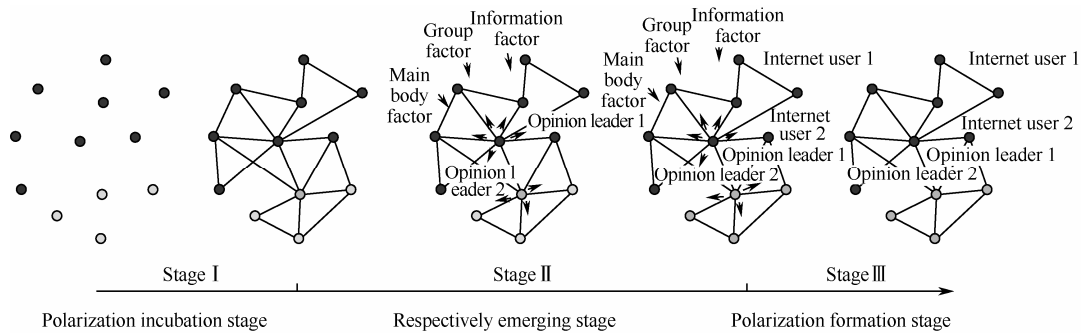


Figure 8-10 The process of polarization formation

Stage I, Latency: since the advent of the Web 2.0 era, self-organising Internet users have become the major force of forming networks and enormous user-generated contents (UGC) have been created. If some UGC involve highly sensitive and public events, which are also explicitly described, they are most likely to initiate group discussions and attract attentions across the Internet.

At first, Internet users are taken as varied and isolated points. Individuals in the same group sharing the same interests are marked with the same color. When they start to follow an event, individual Internet users will interact and discuss with each other; two individuals in interaction or discussion will be connected with edges. This process is the initial stage of group polarization.

Stage II, Emergence: a certain individual with high perceived benefits (for example, an opinion leader) notices the event and posts their personal views on the Internet; at the same time, another individual of the same status (for example, another opinion leader) may post an opposing view on the Internet. At this point, under the guidance of the nonlinear coupling theory, the subject, the group and the topic factors will synthetically affect the influence of other individuals on the judgement of viewpoints. Therefore, when observing the views of the above two individuals, other individuals will choose a point of view to follow under the influence of the polarization effect.

As the polarization progress progresses and more users start to follow a point of view, more and more individuals will believe in the validity of this view; thus, group polarization begins to form gradually. This process is the emergence stage of group polarization.

Stages III, Formation: As can be seen from Figure 8-10, at stage II, Internet user 1 has not interacted with any opinion leaders, so he always maintains his original opinion. In the third stage, however, because his neighbors follow the opinion of opinion leader 1,

under the influence of others, user 1 adaptively aligns his opinion with that of opinion leader 1. As for Internet user 2, we can observe that he is connected to both opinion leader 1 and opinion leader 2, but he is in the same interest group as opinion leader 1 (the two are more homogeneous); therefore, he is more likely to follow the views of opinion leader 1.

Finally, the views in the system gradually reach a stable equilibrium state: polarization takes shape at both ends. At this point, the group polarization is finally formed. This is the third stage of the overall process.

3. Construction of A Model of Group Polarization in Social Networks

In this part, group polarization in social networks is modelled as follows. It is also a multi-agent based modelling method.

Agent Layer: two heterogeneous agent categories are set up, which are respectively defined as InternetUser and OpinionLeader.

Individual Agent Characteristics Model Layer:

Internal status:

- (1) Individual independence, defined as variable “Ind_thinking(I_e) $\in (1,5)$ ”.
- (2) Group diversity, defined as variable “Group_var(G_k) $\in (1,5)$ ”.
- (3) Ratio of the views from different opinion leaders, defined as variables “OLopinion1” and “OLopinion2”.
- (4) Perceived benefits, defined as variable “Percep_benefit”.
- (5) Information sensitivity, defined as variable “Topic_Sensi(T_s) $\in (1,5)$ ”.
- (6) Information publicness, defined as variable “Topic_Pub(T_p) $\in (1,5)$ ”.
- (7) Information ambiguity, defined as variable “Topic_Ambi(T_a) $\in (1,5)$ ”.

Perceptron: defined in this model as the interaction between users and viewing UGC: interact() and viewUGC().

Effector: defined in this model as changing one's own opinion: opinionChange().

Environment: in this model, two output variables are included to reflect the characterization of group polarization, which are the Polarization Time and the Polarization Percentage.

MAS layer: the relevant rules of group polarization in social networks are constructed based on the model of Michael W. Macy and other classic Hopfield network models.

$$P_{acc} = \frac{\sum_{i=1}^{N-1} I_0 D_0}{N-1} \quad (8-12)$$

Where $N = \rho \cdot A$ is the number of Internet users in the system, ρ is the density of the population, and A is the area of the simulation area. When faced with two different views, if an Internet user adopts point 1, then $I_0 = -1$, if an Internet user adopts point 2, then $I_0 = +1$. D_0 represents the number of internet users who shift from their point of view to point 1 or point 2. P_{acc} is the group pressure on an individual user from an individual with who the former has an interaction with, when he or she adopts or rejects a certain point of view.

$$\text{Percep_benefit} = \frac{V_s}{1 + e^{-P_{acc} \cdot \text{Co_Factor}}} + (1 - V_s) \cdot P_{ex} \quad (8-13)$$

$$\text{Co_Factor} = \frac{T_s \cdot T_p}{\lambda \cdot I_e \cdot G_k \cdot T_a} \quad (8-14)$$

In the above equation, $V_s \in [0,1]$ is the parameter that regulates an individual's perceived press ratio from either inside or outside the system. Co_Factor is the comprehensive influencing factors of the group polarization effect, including the subject factor, the group factor and all variables of information factors, which ranges from 0 to 1. The higher value of Co_Factor, the more significant the group polarization effect, where λ is the adjustment parameter and P_{ex} is the perceived press outside the system.. Here, we set a threshold value $\pi_{\text{threshold}} = 0.7 + \beta\chi$, where β is an regulation parameter and χ is a random value at $(-0.5, 0.5)$. When $\text{Percep_benefit} > \pi_{\text{threshold}}$, individual Internet users will change their point of view to view 1 or view 2; otherwise, their maintain their views unchanged.

4. Model Simulation and Results

According to the the above equation, we assume that $V_s = 1$ and that all the perceived press is from within the system; namely, all the information transmissions are from the same medium, coinciding with hypothesis 1. At the same time, we set parameters λ and β to 25 and 0.5, and the simulation time is in seconds, to achieve ideal simulation results.

Table 8-9 shows the parameters in two extreme cases of group polarization, where Case 1 reflects the easiest setting to produce group polarization, and Case 2 reflects the most difficult setting to produce group polarization. The corresponding simulation results in both cases are shown in Figure 8-11 and Figure 8-12 below ^①.

① Reference link: <http://pan.baidu.com/s/1dDh8M2l>.

Table 8-9 Simulation of two extreme cases of group polarization

	Individual Independence	Group Diversity	Information Sensitivity	Public Nature of the Information	Information Ambiguity
Case 1	1	1	5	5	1
Case 2	5	5	1	1	5

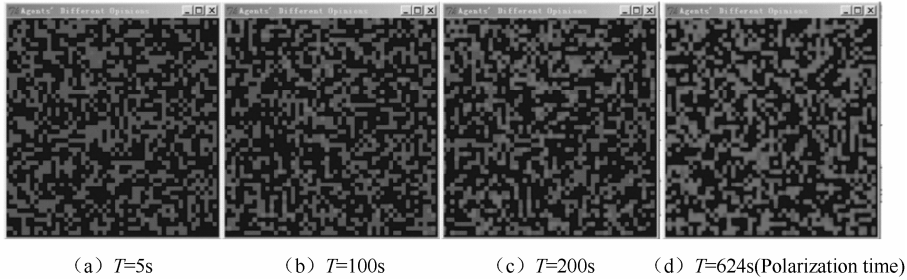


Figure 8-11 Simulation result from Case 1

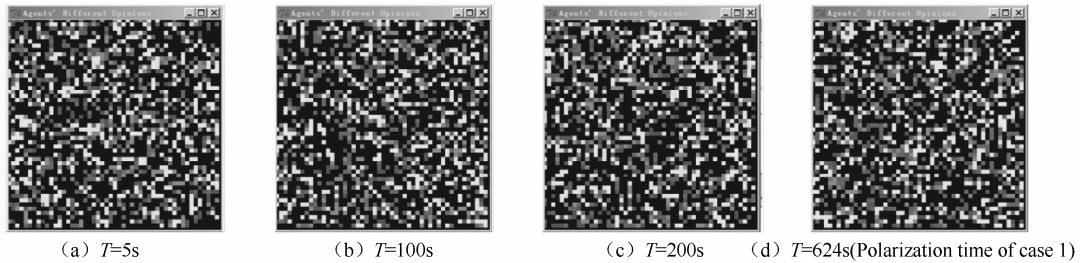


Figure 8-12 Simulation result from Case 2

Figure 8-11 shows that the value of group diversity is 1, indicating that all Internet users belong to the same group, as represented by squares of the same brightness. In the initial stage, each Internet user holds his or her own point of view. After the opinion leaders disseminate their views in the network (represented by squares of two different brightnesses), the Internet users are affected by a combination of factors, and they choose to follow other people's point of view (represented by squares shifting from the initial brightness to the other two brightnesses, as shown in Figure 8-11). Finally, when 80 percent of Internet users have changed their views to form two extremes (indicated by squares in the other two brightnesses), group polarization arises.

Figure 8-12 shows that the group diversity of Internet users is 5 (represented by five kinds of brightness), similar to the evolution of Case 1; however, obviously we can see that

only a few individuals change their views over time in Case 2. When the polarization time in Case 1 ($T=624s$) is reached, polarization does not arise in Case 2 (the brightness indicating Internet users' views remain unchanged). Therefore, this model can effectively simulate the evolution of group polarization.

Then, we changed the value of each influencing factor (individual independence, group diversity, population density, information sensitivity, information publicness, information ambiguity), and we observed the effect of such changes on group polarization. Furthermore, we ranked the influencing factors based on the extent to which each of them influences group polarization, and achieved the following result: group diversity > information ambiguity > individual independence > information sensitivity > information publicness. Finally, we determined the threshold at which all the influencing factors can synthetically influence group polarization. When $Co_Factor > 0.2$, despite the differences in the time and the percentage of group polarization in different cases, polarization will eventually occur. By contrast, when $Co_Factor \leq 0.2$, group polarization basically does not occur. This conclusion proves the critical influence of the integrated influencing factors on sudden events in the system, especially on group polarization. You may refer to the original text for more in-depth and comprehensive information^①.

8.3.7 Simulation of Group Polarization in Social Networks With the Influence of Social Network Structure^②

In the previous section, we did not consider the influence of network structure on group polarization; in this section, we will elaborate on the important influence and significance of the social network structure.

This section first presents an improved local-world evolving model and a group polarization model, and then theoretically describes the impact of social network structure on group polarization. For the empirical study, the social relations between members of a laboratory were examined to verify the correctness of the models, and proposed the rules in which social network structure influences group polarization. Computer simulation methods were used to demonstrate that different platforms have difference influences on information diffusion from the perspectives of topic networks and real networks

① Refer to the following article: Shiyu Du, Jiayin Qi. Multi-agent Modeling and Simulation on Group Polarization Behavior in Web 2.0 [J]. Journal of Networks, 2014.

② Refer to the following article: Shiyu Du, Jiayin Qi. Study on the Influence of Social Network Structure on the Polarization of Group Opinions [J]. Journal of Information Systems, 2014.

respectively. Finally, the following issues were investigated:

- (1) How does the social network structure influence group polarization?
- (2) Do different social network platforms influence group polarization differently?
- (3) For different scenarios, which social network platforms are best suited for diffusing information and reducing group polarization. By answering this question, we can further explore the function mechanisms of the network structure and social media, develop better information diffusion strategies, and improve the effectiveness of information diffusion, which is of certain reference value for improving online business models in social networks and mining the potential application values.

1. Improved Local-world Evolving Model

Based on the improved local-world evolving model, this part of the research is aimed to construct an evolving undirected scale-free network model (EUSN model) and an evolving directed scale-free network model (EDSN model). The EUSN model is mainly applied to typical undirected online social networks, such as Renren and Facebook; friends' relations in this kind of networks are bi-directional. By contrast, the EDSN model is mainly used to simulate typical directed online social networks, such as Sina Weibo and Twitter; friends' relations in this kind of networks are unidirectional. These two models describe the evolution of an online local world in a nonlinear way, and more accurately describe the structural characteristics and evolution of online social network platforms.

The optimum selection probability of the EUSN model is shown in equation (8-15), where $P(i)$ is the probability of a certain node i with a node degree of $k(i)$ being connected. By contrast, for the EUSN model, both indegree k_{i_in} and outdegree k_{i_out} of node i need to be considered, which is shown in equation (8-16) and equation (8-17).

$$P(i) = \frac{k_i^{1+0.5\log_{10} k_i}}{\sum_j k_j^{1+0.5\log_{10} k_j}} \quad (8-15)$$

$$P_{out}(i) = \frac{k_{i_out}^{1+0.5\log_{10} k_{i_out}}}{\sum_j k_{j_out}^{1+0.5\log_{10} k_{j_out}}} \quad (8-16)$$

$$P_{in}(i) = \frac{k_{i_in}^{1+0.5\log_{10} k_{i_in}}}{\sum_j k_{j_in}^{1+0.5\log_{10} k_{j_in}}} \quad (8-17)$$

The evolution process of this network includes the addition of new nodes and the generation and extinction of new links. Moreover, according to relevant theories, users prefer to establish contact with “closely related nodes”; therefore, the specific process is as

follows.

(1) Network initialization: the initial network is a random network with m_0 nodes and e_0 edges, where the connectivity of each node is at least 1, so that the network has no isolated nodes.

(2) Network evolution: in each time interval, $1/M$ of all the nodes are randomly selected as a local network, and the following process is repeated according to a certain probability.

The addition of new nodes: a new node is added to the selected local network with a probability of p_1 , the new node is connected to m_1 nodes according to the optimum selection probability, and the member relationship matrix is updated.

New link generation mechanism I: m_2 new links are added to the network with a probability of p_2 . A node is randomly selected within the network, and one of its neighbors' neighbors is selected as the opposing node, to establish a link. Repeat this action for m_2 times, and update the member relationship matrix.

New link generation mechanism II: m_0 new links are added to the network with a probability of p_3 , and a node is randomly selected within the network. Then, for the EUSN model, the opposing node is selected among nodes in the local world according to equation (8-15), and a link is established between them, generating m_2 undirected links. For the EDSN model, a node is selected according to equation (8-16) as a follower, and a node from the local world is selected according to equation (8-17) as the followed, generating m_2 directed links.

Distinction of new links: one edge of the network is removed with probability p_4 . If the removal of this edge causes an isolated node, then the links are rematched using the above steps. If the network becomes several unconnected small groups, then give up removing the link.

2. Group Polarization Model Based on Social Network Structure

To determine the perceived stress of every Internet user in the network, the PageRank algorithm is used to evaluate the priority for visiting the user (Lei, 2009). According to the member relationship matrix $A_{N \times N}$, the PageRank matrix $R_{1 \times N}$ of the nodes is generated. Moreover, based on social comparison theory, we know that the path distance between two nodes affects the level of "trust" between the members (Latan B., 1981). Therefore, the shortest path between nodes is assumed to be related to the level of perceived stress, and the shortest path matrix between two nodes is represented as $\{D_{ij}\}$.

According to classical theories on group polarization, a user's perceived press from the other nodes is directly proportional to the influence of the relevant nodes (indicated by PageRank value R_j) and inversely proportional to the shortest path between the nodes (indicated by the shortest path matrix $\{D_{ij}\}$). Thus, the representation of perceived stress is shown in equation (8-18), where I_{ij} represents the perceived stress applied by node j within the group influence matrix $I_{N \times N}$ on node i .

$$I_{ij} = R_j / D_{ij} \quad (8-18)$$

In the network, the perceived group stress P_i of each node, which represents the average value of pressure applied by all the members of the network, is evaluated by using equation (8-19), derived from the Hopfield network model. N is the number of nodes in the network, S_j is the opinion value of node j , and when node j holds a supporting, opposing or neutral attitude, $S_j=1, -1$ or 0 , which is captured in the group view matrix $S_{1 \times N}$.

$$P_i = \frac{\sum_{j=1}^N I_{ij} S_j}{N} \quad (8-19)$$

The final opinion of the group members depends not only on the overall group pressure but also on the initial point of view. Therefore, a change in point of view is also related to the value of $\beta \times S_i + (1 - \beta) \times P_i$. S_i is node i 's initial point of view, while β is an adjustable parameter in the $(0, 1)$ range, whose value is set to 0.5 in the simulation process.

3. The Integrated Model

The integrated model is a combination of the local-world evolving model and the group polarization model. As shown in Figure 8-13, it includes three stages of the simulation process.

Stage 1: Network simulation

In this stage, first, the given number of nodes and the number of links of the real network is entered, the corresponding parameters are adjusted, and the corresponding artificial simulation network is obtained. Secondly, by comparing the structure parameters of the artificial simulation network and the real network, we can determine whether the artificial simulation network can describe the structure and information diffusion characteristics of the real network.

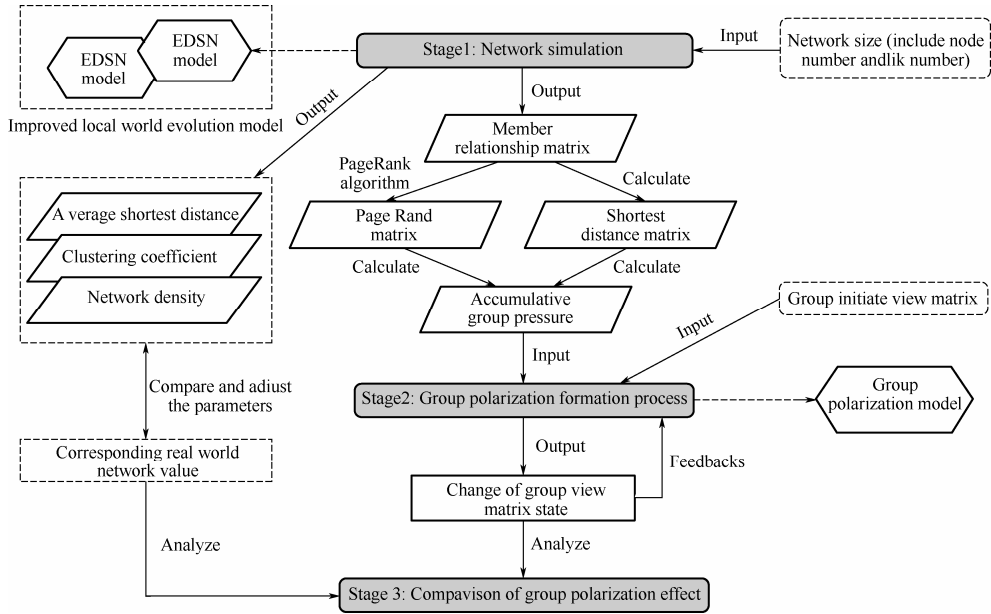


Figure 8-13 Model structure

Stage 2: Process of group polarization formation

According to the group polarization model, threshold π_{thresh} is randomly set. When $\beta \times S_i + (1 - \beta) \times P_i > \pi_{\text{thresh}}$, node i 's view in the initial opinion matrix is updated to a value of 1, and when $\beta \times S_i + (1 - \beta) \times P_i < -\pi_{\text{thresh}}$, node i 's view is updated to a value of -1; otherwise, the value of node i 's view remains unchanged. This process is repeated until a stable state of polarization is reached.

Stage 3: Comparison of the group polarization effects

Two indicators reveal the group polarization effect: time and quantity. Group polarization time refers to the length of time required for the dominant emotion in the group to reach 0.9. The proportion of group polarization refers to the proportion of each view of users after polarization has been reached on the same platform.

On a single platform, while holding the values of all the other variables constant, adjust the corresponding network structure parameters, to determine how the structure of the online social network influences the effect of group polarization. As for different online topic networks and real networks, by comparing the different polarization times and proportions of different platforms, we can analyze whether different social networks have different effects on group polarization, and which kind of social network platforms are most suitable for diffusing information and reducing group polarization in different application scenarios.

4. How Changes in the Structure Parameters of A Social Network Affect Group Polarization

In the study, we selected social network relationships between laboratory members (13 teachers and 108 postgraduates) as the empirical research objects, and then compared and analyzed the structures of four different social networks (Fetion ,QQ, Renren, Sina Weibo) by using the social network analysis (SNS) method, thus to study the impact of social network structure on group polarization.

By comparing the outputs of phase I in the integrated model with the structure parameters of a real network (see Table 8-10), we found that the structure parameters of the artificial simulation network are basically consistent with those of the real network. Moreover, the degree distribution of the simulation network follows a power-law distribution, and it is in accordance with the small-world characteristics of online social networks. Therefore, that the ability of our model to simulate a real social network is verified.

Table 8-10 Network structure parameters of online social network platforms

	Renren		Fetion		QQ		Sina Weibo	
	Simulation value	Real value	Simulation value	Real value	Simulation value	Real value	Simulation value	Real value
Number of nodes	121	121	121	121	121	121	121	121
Number of Links	2108	2099	3880	3879	1692	1677	1626	1623
Average degree of the network	17.4215	16.475	32.066	32.41	13.984	12.443	13.438	12.163
Network density	0.14518	0.1334	0.2672	0.2544	0.1165	0.0776	0.112	0.0996
Average shortest path	1.9573	1.928	1.7331	1.747	2.0497	2.21	3.0267	2.832
Clustering coefficient	0.44707	0.479	0.533	0.581	0.4254	0.47	0.1296	0.1995

After simulating the real social network, we further examined the general rule by which social network structure affects group polarization. Here, we chose to take changing the clustering coefficient in Weibo as an example of simulation. As shown in Figure 8-14, while maintaining other parameters unchanged, we increased the value of the clustering coefficient in Weibo by approximately 50% (the clustering coefficient is 0.2985 on the left side of Figure 8-14 and 0.4590 on the right side of Figure 8-14), we were able to determine the impact of this change on the time and ratio of polarization. When the polarization ratio (shown in Figure 8-14) reaches 0.9, the amount of time before polarization is reduced from 34 to 18. Moreover, as can be seen in Figure 8-15, for the same time steps, the polarization

ratio is significantly improved with an increase in the clustering coefficient. Therefore, it can be concluded that if the other parameters are kept constant, increasing the clustering coefficient can promote the formation of polarization, as it shortens the time before polarization and increases the polarization ratio.

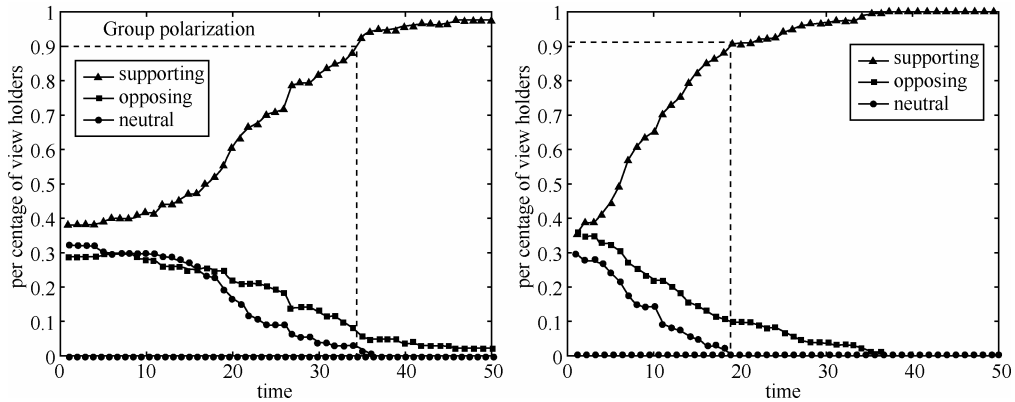


Figure 8-14 Comparison of group polarization times for different clustering coefficients based on Sina Weibo platform data

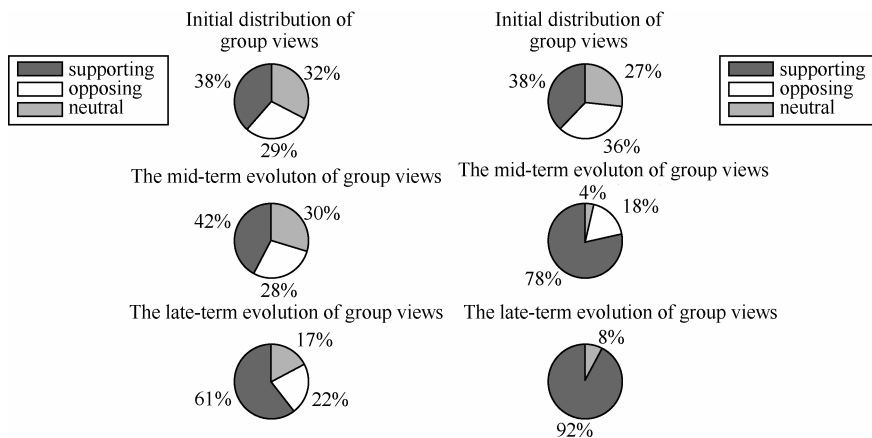


Figure 8-15 Comparison of the proportion of users showing group polarization for different clustering coefficients based on Sina Weibo platform data

Moreover, we also simulated the clustering coefficient, network density and average shortest path for different platforms; however, owing to space limitations, we list only the relevant conclusions here.

- (1) The greater the clustering coefficient, the larger the polarization effect.
- (2) The greater the network density, the larger the polarization effect.

(3) A change in network density will cause a change in the average path length. The greater the network density and the shorter the average path length results are, the larger the polarization effect.

5. Comparison Analysis of Group Polarization Between Different Social Network Platforms

1) Network simulation of different platforms based on topic networks

To expand the scope of the study, we analyzed the similarities and differences in the effects of group polarization and information diffusion in different online topic networks. We first selected the typical case of “Fang Han incident” in 2012 as a study object and then collected data relevant to this case from RenRen and Sina Weibo, including event nodes, connections between nodes, each node’s view on this event and other relevant data.

On the Renren platform, we searched the keyword “Han Fang” and obtained a popular blog with high numbers of views (116,232), shares (22,773) and comments (1039). Considering that the connection information about this user cannot be determined based on whether or not users have “read” or “forwarded” the information, we examined the comments to determine the connections between nodes, and we extract 165 nodes from top ranked comments by reading comments one by one to assess the relationships and emotions in the comments among these nodes. Following the data collection method for the Renren platform, we selected a popular blog from Sina Weibo’s 63,251,307 microblogs on “Han vs. Fang” with high numbers of forwards (28,137) and comments (12,377), and we again extracted 165 nodes to determine the connections between these nodes.

By the adjusted model parameters, we construct two artificial topic networks respectively based on the two platforms, to obtain the graphs of group polarization effects, with which we compared the different effects of different social network platforms on group polarization regarding the same topic. Here, the same analysis method as above is used, as shown in Figure 8-16. We found that when the polarization ratio reaches 0.9, the time before polarization is 42 on Sina Weibo and 35 on Renren. Moreover, as shown in Figure 8-17, with the same time step, the polarization ratio of Sina Weibo is generally greater than that of Renren. The simulation results show that, if viewed from the effects of a particular topic, in the network structure of Sina Weibo is more conducive to the generation of group polarization. However, whether or not this conclusion is applicable to a

real social network requires further study.

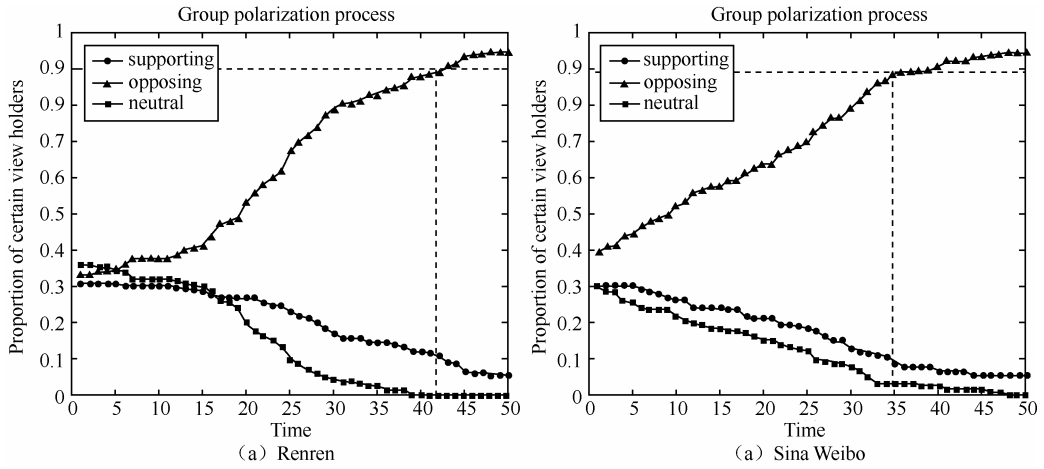


Figure 8-16 Comparison of group polarization times in Renren and Sina Weibo topic networks

2) Network simulation based on different real network platforms

In this study, based on the relevant data about real networks in the reference (Fu et al., 2007; Jin Xin et al., 2012), the group polarization effects of different social network platforms are compared by adjusting the analog network parameters and simulating different real social network platforms described in the reference (see Figure 8-17).

Because of the large size of the real network data, we simplified the step 4 of network evolution (i.e., “the death of new links”) in the model in order to improve the efficiency of the simulation. The real network simulation results based on Renren and Sina Weibo demonstrate that different online social network platforms can exert different group polarization effects. As Figure 8-18 shows, on the artificial network, the time before polarization for Sina Weibo is 36 when the group polarization ratio reaches 0.9; however, as time goes on, the proportion of users showing polarization in Renren does not change substantially, with only an increase of 0.035 percentage points. Further, as Figure 8-19 shows, with the same time step, the polarization ratio of Sina Weibo is far greater than that of Renren, and the polarization phenomenon generally does not appear on the Renren platform. Hence, considering the overall network structure, we can conclude that it is difficult for group polarization to arise on the Renren platform, whereas the Sina Weibo network structure facilitates the spread and diffusion of information.

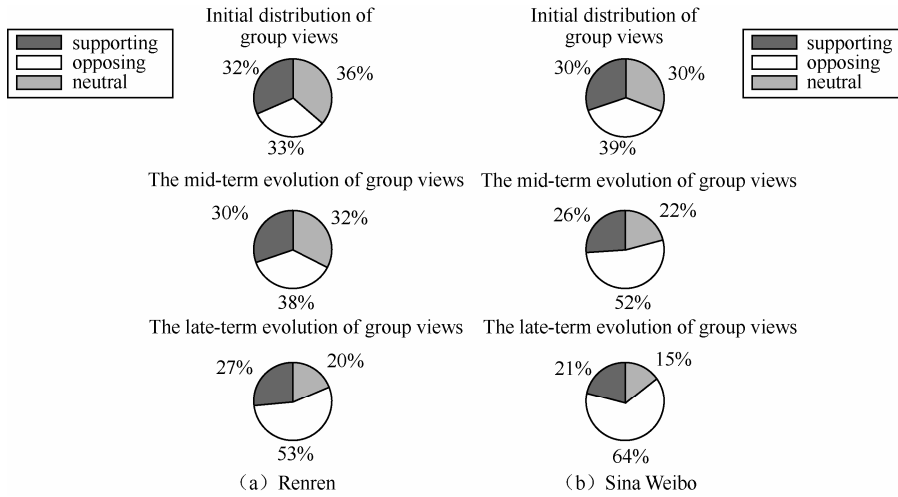


Figure 8-17 Comparison of group polarization ratios in Renren and Sina Weibo topic networks

3) Conclusion

This part of the study explores the impact of social network structure. We found that both clustering coefficient and network density have a positive effect on the formation of group polarization, while the average shortest path has a negative effect on it. These simulation results show that as far as the single-platform social network structure and groups are concerned, a compact and dense network structure fosters the formation of group polarization.

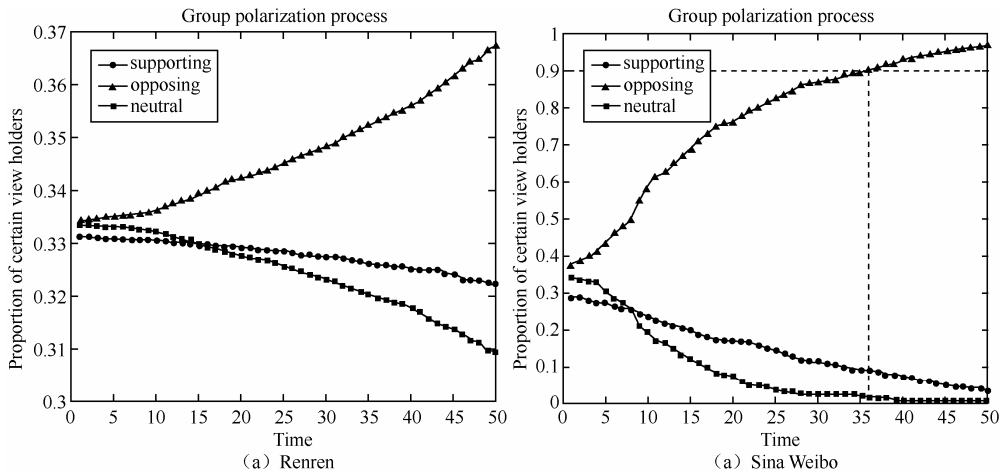


Figure 8-18 Comparison of group polarization time in Renren and Sina Weibo networks

Regarding networks with intense communication, a social network platform based on instant messaging (such as the Fetion network in this study) engenders the most remarkable effect on information diffusion among the four categories of online interactive applications, and it is most conducive to the spread and diffusion of information. Regarding topic-based networks, microblog-type social network platforms (such as Sina Weibo in this study) and online social network platforms (such as Renren in this study) both show significant group polarization, with microblogging applications having the more significant effect. Overall, microblogging social network platforms (such as Sina Weibo in this study) show more significant group polarization, and online social network platforms (such as Renren in this study) generally show no polarization, due to the lack of topic guidance.

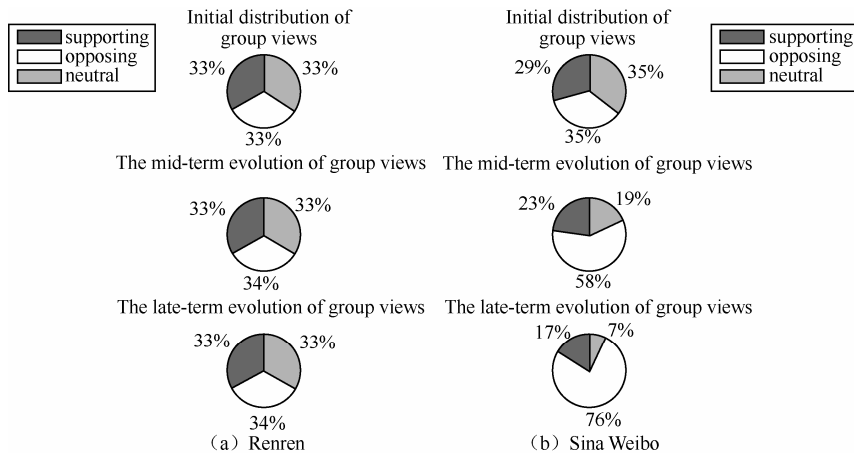


Figure 8-19 Comparison of group polarization ratios in Renren and Sina Weibo real networks

8.4 Summary

This chapter examines the mechanism of two types of group behaviors, namely, collective intelligence and group polarization, in order to explain group aggregation and the mechanisms that influence it. With respect to group intelligence, we first described the specific concepts of group intelligence and the related theories, including self-determination theory. We then discussed the status quo of researches in this area from the perspectives of the conditions and factors that engender collective intelligence and group polarization, as well as analytical models that describe them. Finally, we concluded the research results of our study. Finally, we elaborated on group polarization in the similar structure.

The evolution of human civilization is realized by individuals acting in groups, and gathering in groups is the only feasible way to win victories. Because of the various mechanisms that influence group behavior, the organization and ideology of groups tend to be diversified. Although great theoretical and practical achievements have been made in the research of group behavior, research on group intelligence and group polarization, especially those on the Internet, is still a nascent field, allowing further thought and development.

(1) The description of group intelligence is still considered to be in a “contending” state ^[59]. While most studies recognize the concept of “group intelligence”, we cannot ignore some of the dialectical vision with respect to the concept. For example, Charles Mackey scoffed at the idea that “a group knows everything”; in his view, group judgment is bound to be extreme. The world’s investment guru Bernard Baruch (Bernard Baruch, 1870—1965) even remarked that anybody who is considered a very sensitive and rational individual will soon become an idiot when he or she joins a group. Group intelligence has a long history of development over human history. It is not the development of group intelligence that will inevitably lead to the development of a pluralistic society but the development of a pluralistic society that will promote the development of group intelligence. However, most scholars now believe that a group is not merely a mob; rather, it has its own mobilization, decisions, actions and logic of risk aversion, which create certain conditions that will generate group intelligence. As the ancient Chinese stated, “Three cobblers with their wits combined equal Zhuge Liang, the mastermind”; thus, perhaps group intelligence does exist, and furthermore, it will play an increasingly important role in the future development of human society. For example, in studying the speakers and audiences of speech, Ramine Tinati et al. ^[60] demonstrated group intelligence in the application of “Online Citizen Science”. Moreover, Tony Diggle et al. ^[61] showed how group intelligence can help in finding a solution to the global scarcity of water. These are issues of great practical significance that are worth further exploration. In addition, our next research focus is the application of group intelligence to solve real-world problems.

(2) Regarding group polarization, if guided improperly, it would undoubtedly have a negative impact on society, and the failure to take effective measures will lead to extreme behavior that may threaten the normal order. Although sufficient evidence demonstrates the prevalence of group polarization, the way in which polarization arises and its effects are still worth research attention. For example, in a study of group polarization, Fishkin and Luskin found that group polarization did not arise in many groups even though they showed

behavioral consistency. Similarly, not all groups polarize in the same way. Moreover, under certain conditions, strong group polarization can also have a very positive effect. For example, group polarization can consolidate a dominant ideology and prevent the social conflicts from breaking the “safety valve” (Song Jiageng, 2010). Companies may also be able to take advantage of group polarization in promoting their products and forming an invisible force. Other research shows that well-established groups are less likely to generate group polarization, since the main purpose of these groups is to solve problems and group members are well informed of the content of the problems (Shi Bo, 2010). However, when a group is newly established or faced with new tasks, group polarization can have a more profound impact on the process of group decision making. In recent years, research on group polarization has continually progressed. For example, Muste Christopher et al. ^[62] redefined polarization as the differentiation and “culture war” of social groups. Further, David H. Zhu et al. ^[63] studied the influence of polarization on the workshop decision making of businesses or companies. Clearly, how group polarization evolves, what effects it engenders, and how group polarization theory can be integrated with practical issues constitute the major research areas in the future.

References

- [1] Gustave Le Bon. The crowd: A study of the popular mind [M]. Macmillan, 1897.
- [2] McLuhan Marshall. Understanding media: The extensions of man [M]. MIT press, 1994.
- [3] David W. McMillan David M. Chavis. Sense of community: A definition and theory [J]. Journal of community psychology, 1986, 14(1): 6-23.
- [4] Marvin E. Shaw. Group dynamics: The psychology of small group behavior[J]. 1971.
- [5] Marcello Andrea Canuto, and Jason Yaeger, eds. The archaeology of communities: A new world perspective. Psychology Press, 2000.
- [6] Iain Pears. An instance of the fingerpost[M]. Random House, 1998.
- [7] William Morton Wheeler. Ants collected in British Guiana by the expedition of the American Museum of Natural History during 1911[J]. Bulletin of the American Museum of Natural History, 1916, 35: 1-14.
- [8] Émile Durkheim. Rules of sociological method[M]. Simon and Schuster, 1982.
- [9] Martijn C. Schut. On model design for simulation of collective intelligence[J]. Information Sciences, 2010, 180(1): 132-155.
- [10] Edward L. Deci. Intrinsic motivation[M]. New York: Plenum Publishing, 1975.
- [11] James Surowiecki. The wisdom of crowds[M]. Random House LLC, 2005.

- [12] Hayes Tom, Michael S. Malone. No Size Fits All: From Mass Marketing to Mass Handselling [M]. Penguin, 2009.
- [13] Don Tapscott, Abthony D. Williams. Wikinomics: How mass collaboration changes everything [M]. Penguin, 2008.
- [14] Thomas W. Malone, Klein Mark. Harnessing collective intelligence to address global climate change [J]. *innovations*, 2007, 2(3): 15-26.
- [15] Sinan Aral, Dylan Walker. Identifying influential and susceptible members of social networks [J]. *Science*, 2012, 337(6092): 337-341.
- [16] Simmel Georg. Sociological theory [J]. 2008.
- [17] Ioanna Lykourantzou, Papadaki Katerina, Dimitrios J. Vergados, et al. CorpWiki: A self-regulating wiki to promote corporate collective intelligence through expert peer matching[J]. *Information Sciences*, 2010, 180(1): 18-38.
- [18] Wu-Chih Hu. Deriving collective intelligence from reviews on the social Web using a supervised learning approach[J]. *Expert Systems with Applications*, 2011, 38(10): 13149-13157.
- [19] Thomas D. Seeley. The wisdom of the hive: the social physiology of honey bee colonies [M]. Harvard University Press, 2009.
- [20] Scott E. Page. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition) [M]. Princeton University Press, 2008.
- [21] Everett Stiles, Xiaohui Cui. Workings of collective intelligence within open source communities [M]//*Advances in Social Computing*. Springer Berlin Heidelberg, 2010: 282-289.
- [22] Anna De Liddo, Simon Buckingham Shum. Cohere: A prototype for contested collective intelligence [J]. 2010.
- [23] Anna De Liddo, Sándor Á, Buckingham Shum Simon. Contested collective intelligence: Rationale, technologies, and a human-machine annotation study [J]. *Computer Supported Cooperative Work (CSCW)*, 2012, 21(4-5): 417-448.
- [24] Elisabeth Noelle-Neumann. The spiral of silence: Public opinion--Our social skin [M]. University of Chicago Press, 1993.
- [25] Mohamed, A. Amin, Frank A. Wiebe. Toward a process theory of groupthink[J]. *Small group research*, 1996, 27(3): 416-430.
- [26] Cartwright, Dorwin. The nature of group cohesiveness[J]. *Group dynamics: Research and theory*, 1968, 91: 109.
- [27] Mullen, Brian, Copper Carolyn. The relation between group cohesiveness and performance: an integration [J]. *Psychological bulletin*, 1994, 115(2): 210.

- [28] Liu Hong. The group decision in the menu[J]. Business: review. 2012 (4): 106-108.
- [29] Wangchao Hui, Luo Xinxing. A study on the relationship between intellectual capital and different types of innovation[J]. East China Economic Management. 2010 (1): 109-114.
- [30] John C. Harsanyi. Games with Incomplete Information Played by “Bayesian” Players, I-III Part I. The Basic Model[J]. Management science, 1967, 14(3): 159-182.
- [31] Sun Yuqiu, Chen Shengtao. Bayesian method for forecast of stock. Journal of Guangdong polytechnic Normal University. 2003 (4): 78-80.
- [32] Eric Bonabeau, Marco Dorigo, Guy Theraulaz. Swarm Intelligence: From Natural to Artificial Systems[M]. New York: Oxford University Press, 1999.
- [33] Marco Dorigo, Luca Maria Gambardella. Ant colony system: a cooperative learning approach to the traveling salesman problem [J]. Evolutionary Computation, IEEE Transactions on, 1997, 1(1): 53-66.
- [34] Wang Yanlin, Li Longshu, Hu Zhe. Swarm Intelligence Optimization Algorithm[J]. Computer Technology and Development. 2008, 18(8): 114-117.
- [35] James F. Kennedy, Russell C. Eberhart. Swarm intelligence[M]. Morgan Kaufmann, 2001.
- [36] Yang Wei, Li Qiqiang. Study of the Variable Load While Estimating Existing Bridge Structure[J]. Engineering Science. 2004, 6(5): 87-94.
- [37] Russell C. Eberhart, Xiaohui Hu. Human tremor analysis using particle swarm optimization[C]// Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on. IEEE, 1999, 3.
- [38] Stoner James Arthur Finch. A comparison of individual and group decisions involving risk[D]. Massachusetts Institute of Technology, 1961.
- [39] Scharfstein David S, Jeremy C. Stein. Herd behavior and investment[J]. The American Economic Review, 1990: 465-479.
- [40] Bikhchandani Sushil, Hirshleifer David, Welch Ivo. A theory of fads, fashion, custom, and cultural change as informational cascades [J]. Journal of political Economy, 1992: 992-1026.
- [41] Leon Festinger. A theory of social comparison processes [J]. Human relations, 1954, 7(2): 117-140.
- [42] Mackie Diane M. Social identification effects in group polarization [J]. Journal of Personality and Social Psychology, 1986, 50(4): 720.
- [43] Tajfel Henri. Experiments in intergroup discrimination[J]. Scientific American, 1970, 223(5): 96-102.
- [44] Tajfel Henri, Billig M G, Bundy R P, et al. Social categorization and intergroup behaviour[J]. European journal of social psychology, 1971, 1(2): 149-178.
- [45] Van Swol, Lyn M. Extreme members and group polarization[J]. Social Influence, 2009, 4(3):

185-199.

- [46] John Turner, Brown R J, Tajfel Henri. Social comparison and group interest in ingroup favouritism [J]. *European Journal of Social Psychology*, 1979, 9(2): 187-204.
- [47] Cialdini Robert B. *Influence: Science and practice*[M]. Boston, MA: Allyn and Bacon, 2001.
- [48] Shi Bo. A Study of the Dynamic Mechanism and Coping Strategies for Group Polarization of Network Public Sentiment[J]. *Journal of Intelligence*. 2010, 29(7): 50-53.
- [49] Xie Xinzhou. Experiments of Postulate of "Spiral of silence" on Internet[J]. *MODERN COMMUNICATION*. 2004 (6): 17-22.
- [50] Sunstein Cass R. The law of group polarization[J]. *Journal of political philosophy*, 2002, 10(2): 175-195.
- [51] Morris E Charles, Anle Tieu, Kingsley Dixon. Seed coat dormancy in two species of *Grevillea* (Proteaceae)[J]. *Annals of Botany*, 2000, 86(4): 771-775.
- [52] Hong Harrison, Jeffrey D. Kubik, Jeremy C. Stein. Thy neighbor's portfolio: Word - of - mouth effects in the holdings and trades of money managers[J]. *The Journal of Finance*, 2005, 60(6): 2801-2824.
- [53] YANG Shanlin, ZHU Keyu, FU Chao, LU Guangyan. Simulation of the group decision conformity based on cellular automata model[J]. *Systems Engineering Theory & Practice*. 2009 (9): 115-124.
- [54] Watts Duncan J, Peter Sheridan Dodds. Influentials, networks, and public opinion formation[J]. *Journal of consumer research*, 2007, 34(4): 441-458.
- [55] Kwak Haewoon, Lee Changhyun, Park Hosung, et al. What is Twitter, a social network or a news media?[C]//*Proceedings of the 19th international conference on World wide web*. ACM, 2010: 591-600.
- [56] Zhang Yiwen, Qi Jiayin, Fang Binxing. Online Public Opinion Risk Warning Based on Bayesian Network Modeling[J] *Library and Information Service*. 2012, 56(2): 76-81.
- [57] Abhijit V Banerjee. A simple model of herd behavior[J]. *The Quarterly Journal of Economics*, 1992, 107(3): 797-817.
- [58] Meng Chang. Information Cascade and Coordination Failure—A study of a theory based on network extemality[J]. *Journal of Beijing Technology and Business University(Social Science.)* . 2006, 21(5): 93-99.
- [59] Andrew Keen. *The Cult of the Amateur: How blogs, MySpace, YouTube, and the rest of today's user-generated media are destroying our economy, our culture, and our values* [M]. Random House LLC, 2008.
- [60] Ramine Tinati, Elena Simperl, Markus Luczak-Röesch, et al. Collective intelligence in citizen

science—a study of performers and talkers[J]. 2014.

- [61] Tony Diggle. Water: how collective intelligence initiatives can address this challenge[J]. *Foresight*, 2013, 15(5): 342-353.
- [62] Muste Christopher P. Reframing Polarization: Social Groups and “Culture Wars”[J]. *PS: Political Science & Politics*, 2014, 47(2): 432-442.
- [63] David H. Zhu. Group Polarization on Corporate Boards: Theory and Evidence on Board Decisions About Acquisition Premiums[J]. *Strategic Management Journal*, 2013, 34(7): 800-822.

Chapter 9

Information Retrieval in Social Networks

Information Retrieval (IR) is a process of retrieving information, which satisfies users' need, from massive unstructured data (such as natural language texts) sets, which is an important tool that helps users rapidly and effectively acquire useful information from massive data. With the drastic increase of data size and increasingly growing user needs in search services, IR has evolved from a tool only designed library in the beginning into a network service indispensable for life, work, and study. Apart from the search systems represented by the popular Google search engine, some other common forms of IR systems include classification systems, recommendation systems, and Q&A systems.

With the rapid popularization and continual development of social networking services (SNS), IR are not only afforded with new resources and opportunities, but also confronted with new problems and challenges. Acquiring information from such emerging resources as social networks has gradually drawn extensive attentions from both industrial and academic circles. Compared with traditional webpages, social network texts have many different characteristics, such as the limit of text length, the special expression form (such as Hashtag^① in microblogs), and the existence of social relations between authors, etc. These differences make it inappropriate to directly apply traditional IR technologies to an SNS environment. Social network oriented IR technology is still confronted with many problems and difficulties, and it is of great academic significance and application value to conduct research in this field.

① Hashtag here refers to the tag bracketed with “#” in the text of microblogs, also called theme tag, which can be regarded as a mark on a microblog made by the author. After development and evolution, it has been used by some social network sites to represent topics.

This chapter mainly introduces IR for social networks, aimed at giving readers a picture of the challenges faced by IR technology in its applications in the new resources of social networks, and introducing the possible solutions to such problems. Concretely, three most representative IR applications - search, classification and recommendation - are discussed in this chapter. This chapter is arranged as follows: section 9.1 is the Introduction, which mainly introduces the relevant common concepts used in the chapter, and the challenges in front of social network oriented IR technology; sections 9.2, 9.3 and 9.4 respectively introduce the basic methods for content search, content classification and recommendation in social networks, and the status quo of researches in these areas; section 9.5 gives the summary of this chapter and future prospects. Specially, in sections 9.2 and 9.3, the relevant researches are introduced based on data from microblog - one of the most representative SNSs, while section 9.4 is focused on social networks developed from traditional E-commerce websites, which carry social networking information, where commodities are recommended by integrating such online social networking information.

9.1 Introduction

IR is a process of retrieving information (generally documents), which satisfies users' need for information, from massive unstructured data (generally texts) sets (mostly stored in computers)^[1].

The unstructured data refers to data without obvious structural markers, in contrast with structured traditional databases. Natural language texts are the most common unstructured data. The user information need refers to an information theme that the user wants to find, while a query is usually a piece of text that the user submits to the retrieval system which represents his information need, or an object in any other forms (such as one or several keywords in the text search engine or sample images in the image search engine).

The data set being retrieved refers to a corpus or a collection. Each record in a collection is called a document. A document is the basic object to be retrieved, such as a microblog message, a webpage, an image, or a sentence. It should be noted that the document here is different from a file. A file may contain a number of documents, and an IR document may be composed of several files.

During processing, a document is often converted into a representation that can describe the key characteristics of its essential content, and such a characteristic is referred to as a "term" in IR, which is a basic retrieval unit in the retrieval system. Terms in texts

are generally expressed by keywords. For example, in the sentence “social network oriented IR is is very important”, the terms can be “IR”, “social”, “networks”, “very” and “important”. Certainly, in a practical application, the final selected terms depend on the application itself.

The aim of IR is to return and search relevant documents from the document set, and the degree of relatedness is called relevance between a document and the query. As far as IR systems are concerned, it is usually necessary to rank documents based on the relevance between them and the query. In order to overcome the problem that an original query cannot precisely represent the user need, the original query can be modified, which is referred to as query expansion or query reformulation. After the retrieval results are returned, either the user or the system can apply explicit or implicit marks to some documents returned, so as to judge whether they are relevant or not. The original query can be modified based on the marking results. This process is called relevance feedback. If we directly assume that the top k documents of the returned results are relevant and perform feedback based on this assumption, this kind of feedback is referred to as pseudo relevance feedback, and the top k documents are referred to as pseudo relevant documents.

Evaluation is one of the important tasks in IR. Comparing the results returned by the system with the actual results, we can get some evaluation metrics to measure the retrieval results. The most basic evaluation indicators for IR are “precision” and “recall”, with the former referring to the proportion of actually relevant documents in the returned results, and the latter the proportion of actually relevant documents that are returned. The two are respectively used to measure the correctness of the results returned and the degree of coverage of returned results on all correct results.

Example 9-1 Calculation of precision and recall. Provided that there are 100 documents relevant to a query, and a system returns 120 documents, in which 80 documents are actually relevant to the query, the precision of the system in terms of this query is $80/120=2/3$, and the recall is $80/100=4/5$.

Precision and recall are applicable for search and classification tasks in IR, and the dedicated evaluation metrics for recommendation are introduced in section 9.4. Precision and recall are extensively applied, but, in terms of search, the order of returned results which is of vital importance to user experiences is not considered during calculation of precision and recall. Therefore, in the application of information search, the following metrics which take into consideration the order of returned results are usually used: $P@k$ and MAP (Mean Average Precision).

Definition 9-1 $P@k$ (precision at k) refers to the precision of the top k results in the retrieval results; for example, $P@5$ and $P@10$ respectively refer to the ratio of relevant documents to the top 5 results and that to the top 10 results. For a given query q , the $P@k$ is calculated based on the following equation:

$$P@k = \text{Number of relevant documents in the top } k \text{ results} / k \quad (9-1)$$

Definition 9-2 Average Precision (AP) is, given a query q , the average value of the precisions at the positions of all relevant document in the returned result. AP is calculated based on the following equation:

$$AP(q) = \frac{1}{|\text{rel}(q)|} \sum_{i=1}^{|\text{ret}(q)|} \text{isrel}(d_i) \times P@i \quad (9-2)$$

Where $\text{rel}(q)$ is the document collections actually relevant to q ; $\text{ret}(q)$ represents all the returned document collection for q ; d_i is the i th document in the documents returned; $\text{isrel}(d_i)$ is a Boolean function. If d_i is relevant to q , 1 will be returned; otherwise, 0 will be returned.

Example 9-2 Calculation of AP. Provided that the size of the document collection ($\text{rel}(q)$) relevant to a query is 5, in which 4 documents respectively appear on positions 1, 4, 5, and 10 of the search result, the AP of the query will be:

$$(1/1 + 2/4 + 3/5 + 4/10 + 0)/5 = 0.5$$

Obviously, the more in number and the more higher the relevant documents in the returned results, the larger the AP.

Definition 9-3 Mean average precision (MAP) refers to the mean of the AP values of multiple queries. MAP is used to evaluate the quality of an IR system.

Different from traditional IR in many aspects, social network oriented IR has its own characteristics, which have brought both challenges and opportunities for the traditional IR technology:

(1) SNS documents are normally very short, and the contents are sparsely distributed to some extent, which makes it hard to calculate or accurately calculate the similarity due to the scarcity of co-occurrence terms.

(2) The expressions of SNS documents are usually non-standard, streaming and dynamic. New terms mushroom from social networks, which are seriously colloquial. The SNS contents of the same theme may shift over time, so the original expressions and calculation models have to deal with such a shift. This problem is particularly serious for classification.

(3) SNS documents have their own structures and interaction characteristics. SNS documents have their specific ways of expression ways and structures. For example, a microblog document is likely to contain a Hashtag and an external URL link. SNS documents generally contain information about the authors, while social relationships between authors can be formed by way of following or interaction. These characteristics can be used during IR.

(4) The time attribute can be found in most SNS queries and documents, for example, queries are always closely related to the current hot social events, and documents present different distribution characteristics with the passage of time. Using the time attribute to increase the SNS information retrieval effect is an important research direction.

The remaining parts of this chapter will introduce IR in social networks based on three main IR applications or tasks. IR involves an extremely wide coverage and is correlated with other fields. It is impossible to introduce all fields in this chapter due to the length limitation. Readers interested in those topics can refer to social network related papers published in related academic conferences (such as SIGIR and CIKM) in the IR field.

9.2 Content Search in Social Network

A content search task refers to a process of returning relevant information contents from a large amount of information in response to a given query. Content search is one of the most classical application forms of IR. In SNS, content search is urgently needed. For example, a user enters “missing of MH370”, with the aim of getting information about this event. Sometimes it is also possible to realize “Expert Location” based on social networks. For example, if you search “machine learning”, some information about the relevant experts can also be retrieved on SNS. This is a specific search application in SNS. Due to the limited length, this topic is not discussed in this chapter. Readers interested in it can refer to Chapter 8 of the book cited^[2]. The Text Retrieval Conference (TREC) added the sub-task of microblog search in 2011, to promote SNS search, particularly microblog search, by providing standard queries and annotation data collections (Twitter).

The basic process of a traditional content search is as follows: Massive document data constitutes a corpus for the search; the user creates a query that can represent his information need; the query and documents are respectively processed and converted into certain representations; the relevance is calculated using the IR model; the documents are returned to the user in an descending order (based on the calculation). According to the

above process, we can see that in the information retrieval process, the processed objects (documents and queries) should be converted into certain forms of representations first, and then the relevance between the objects should be calculated. The conversion of the retrieved objects into expressions, and the calculation of the relevance between them fall into the scope of IR models. There are currently three classical IR models, including the vector space model, the probabilistic model and the statistical language models. The following part briefly introduces these models and the corresponding feedback models.

9.2.1 Classical IR and Relevance Feedback Models

1. Vector Space Model

In the 1950s, the idea of converting texts into term vectors that bear weight information was put forward^[3], and this idea is exactly the essence of the Vector Space Model (VSM). The modern vector space model^[4] put forward by Gerard Salton (1927—1995) et al. is one of the most extensively used retrieval models in the IR field in the past few decades.

The basic idea of VSM: a query is considered as a document; every document is converted into a vector in the same space; the similarity between all vectors is used to measure the relevance between the query and documents. Each dimension of the vector corresponds to a term, and its value represents the importance of the term in the document. This importance is referred to as weight, which is generally calculated using the TFIDF scheme: TF refers to the term frequency; namely, the number of times that a term appears in a document, denoting the representativeness of the term in the document; DF is the number of documents in the document set that contains the term, referred to as document frequency. DF is generally converted into inverse DF (IDF) for calculation. The IDF of a term (t) is generally calculated based on the following equation^①.

$$\text{IDF}_t = \log \frac{N}{\text{DF}_t} \quad (9-3)$$

Where N is the number of documents in the document set; IDF is the ability of the term to distinguish documents. The IDF of commonly used words, such as “of”, “a” and “the”, is small. That is to say, their ability to distinguish documents is very limited. In the VSM, the

① For the purpose of unification, the log in this chapter refers to logarithm with 10 as the base.

weight of a term is the product of TF and IDF.

Example 9-3 Calculation of TFIDF. Assume that the TF of a term in a document is 2, and that it can be found in 10 out of the 100 documents, the TFIDF of the term in the document will be:

$$2 \times \log(100/10) = 2$$

Similarly, the weight of the query and other terms in the document can be calculated, so as to obtain the vector representations of the query and each document. The similarity is calculated at last. For the VSM, the cosine similarity between vectors is used to calculate the relevance between a query and a document. The cosine similarity between a query (q) and a document (d) is calculated based on the following equation:

$$RSV(d, q) = \frac{\sum_t TF_{t,d} \times IDF_t \times TF_{t,q} \times IDF_t}{\sqrt{\sum_t (TF_{t,q} \times IDF_t)^2} \sqrt{\sum_t (TF_{t,d} \times IDF_t)^2}} \quad (9-4)$$

Where $TF_{t,d}$ is the occurrence frequency (number of times) of the term (t) in the document (d).

Example 9-4: Cosine similarity calculation. Assume that the vector representation of a query is $\langle 2, 0, 1, 0 \rangle$, and that of a document is $\langle 1, 2, 2, 0 \rangle$, the cosine similarity between them will be:

$$RSV(d, q) = \frac{2 \times 1 + 0 \times 2 + 1 \times 2 + 0 \times 0}{\sqrt{2^2 + 0^2 + 1^2 + 0^2} \times \sqrt{1^2 + 2^2 + 2^2 + 0^2}} = \frac{4}{\sqrt{45}}$$

In practice, there are many transformed calculation methods for TF and IDF. In addition, the document length is also considered in some VSM weighting representations. Details of the above methods are described in references [5] and [6].

The VSM is very simple and intuitive, and has good practical results. The idea of vector representation is widely used in various fields. Its shortcoming is the “term independence assumption”, i.e., terms in different dimensions are independent of each other, while in practice this assumption obviously does not hold. For example, in an article that contains the term “Yao Ming”, the probability that “basketball” appears in the article obviously increases.

2. Probabilistic Retrieval Model and BM25 Formula

The probabilistic retrieval model (PRM) was firstly put forward in 1960^[7]. It has developed from theoretical research to practical applications in the past decades. The

PRM-based OKAPI^① retrieval system^[8] has made excellent achievements for many times at TREC conferences. INQUERY^[9], another probabilistic retrieval system, also has good reputations. This section firstly introduces the classical binary independence retrieval model (BIR model) and secondly describes the BM25 formula used in the OKAPI system that is evolved from the BIR model. Probabilistic retrieval models are models of a category, so there are many other probabilistic retrieval models in addition to those introduced herein. The readers can refer to the reference^[10] for more information about probabilistic retrieval models.

In the PRM, the relevance between a query and a document is measured by the probability that the document is relevant to the query. Formally, the model introduces three random variables, i.e., D , Q and R , in which D and Q respectively refer to the document and query, and R is a binary random variable, with the value being 1 or 0 (1 denotes that D is relevant to Q , and 0 otherwise). For a given query $Q=q$, when the document $D=d$, the probabilistic model is used to calculate the probability that the document is relevant to the query.

In the BIR model, the probability $P(R=1|D, Q)$ is calculated using the Bayes formula. In IR, the relevance between a query Q and each document is calculated, so for the same query, we denote the $P(R=1|D, Q)$ with $P(R=1|D)$, and get the following:

$$P(R=1|D) = \frac{P(D|R=1)P(R=1)}{P(D)} \quad (9-5)$$

For the convenience of calculation, the BIR model uses the following log-odds to sort the documents^②:

$$\log \frac{P(R=1|D)}{P(R=0|D)} = \log \frac{P(D|R=1)P(R=1)}{P(D|R=0)P(R=0)} \propto \log \frac{P(D|R=1)}{P(D|R=0)} \quad (9-6)$$

Where $P(D|R=1)$ and $P(D|R=0)$ denote the probability of generating the document D respectively under the condition that $R=1$ (relevant) and $R=0$ (irrelevant). \propto denotes order preserving, i.e., the order of the expression before \propto is the same as that after \propto . In the BIR model, provided that the document D is based on the term collection $\{t_i|1 \leq i \leq M\}$ and is generated according to the multivariate Bernoulli distribution^③ (wherein, M is the number of terms). Consequently, the above formula is transformed into:

① <http://www.soi.city.ac.uk/~andym/OKAPI-PACK/index.html>.

② Obviously, for the two documents D_1 and D_2 , if $P(R=1|D_1) > P(R=1|D_2)$, will hold. That is to say, the log-odds function is order-preserving.

③ Multivariate Bernoulli distribution can be considered as the process of flipping M coins, and each coin corresponds to a term. All terms that all upturned coins correspond to constitute the document D .

$$\log \frac{P(D|R=1)}{P(D|R=0)} = \log \frac{\prod_i p_i^{e_i} (1-p_i)^{1-e_i}}{\prod_i q_i^{e_i} (1-q_i)^{1-e_i}} = \sum_i \log \left(\frac{p_i}{q_i} \right)^{e_i} \left(\frac{1-p_i}{1-q_i} \right)^{1-e_i} \quad (9-7)$$

Where p_i and q_i are the probabilities of occurrence of the term t_i in relevant and irrelevant documents respectively. e_i is an variable with the value being 0 or 1 (if $t_i \in D$, $e_i=1$; otherwise, $e_i=0$), and it denotes whether the term t_i exists in the document D . Parameters p_i and q_i can be estimated, and then the rank of each document can be obtained.

Example 9-5 BIR model calculation. Provided that the query is “Information Retrieval Textbook”, a document D is “Retrieval Courseware”, and the number of terms (M) is 5, the parameters p_i and q_i are shown in Table 9-1.

Table 9-1 A calculation example of BIR model

Term	Information	Retrieval	Textbook	Tutorial	Courseware
$R=1, p_i$	0.8	0.9	0.3	0.32	0.15
$R=0, q_i$	0.3	0.1	0.35	0.33	0.10

Consequently,

$$P(D|R=1) = (1-0.8) \times 0.9 \times (1-0.3) \times (1-0.32) \times 0.15$$

$$P(D|R=0) = (1-0.3) \times 0.1 \times (1-0.35) \times (1-0.33) \times 0.10$$

$$\log(P(D|R=1)/P(D|R=0)) = 0.624$$

The basic BIR model does not consider important factors such as TF and document length. Stephen Robertson et al. have made improvements, and put forward the well-known BM25 retrieval formula^[8] as follows:

$$\text{RSV}(d, q) = \sum_{t \in q} \ln \frac{N - \text{DF}_t + 0.5}{\text{DF}_t + 0.5} \times \frac{(k_1 + 1) \text{TF}_{t,d}}{k_1((1-b) + b \frac{\text{dl}}{\text{avdl}}) + \text{TF}_{t,d}} \times \frac{(k_3 + 1) \text{TF}_{t,q}}{k_3 + \text{TF}_{t,q}} \quad (9-8)$$

Where dl is the length of the document; avdl is the average length of documents in the document collection; $\text{TF}_{t,d}$ and $\text{TF}_{t,q}$ are respectively the term frequency of the term in the document and query; b , k_1 and k_3 are empirical parameters. The formula can also be considered as a calculation formula of the inner product of vectors using different TF and IDF calculation methods.

The advantages of probabilistic models are that they are derived based on the probability theory, and they are more interpretable than VSMs. However, in a calculation using probabilistic models, the assumption of term independence still exists. Moreover, the

parameters in the models need to be precisely estimated.

3. SLM-based Retrieval Models and Query Likelihood Models

The statistical language modeling (SLM) technology attempts to build models for natural languages based on statistics and probability theory, so as to obtain the law and features of natural languages and solve specific problems on language information processing. SLM technology can date back to the early 20th century. At that time, the original intentions were studying the Russian reference and building models for sequences of Russian letters^[11]. Since the 1980s, SLM has been widely applied in such fields as speech recognition, optical character recognition, and machine translation, and has become one of the mainstream technologies for language information processing^[12].

Definition 9-4 Statistical language model. A language is essentially the result of certain probability distribution on its alphabet, which shows the possibility that any sequence of letters becomes a sentence (or any other language unit) of the language. The probability distribution is the statistical language model of the language. For any term sequence $S = w_1w_2\dots w_n$ in a language, its probability can be figured out based on the following equation:

$$P(S) = \prod_{i=1}^n P(w_i | w_1w_2\dots w_{i-1}) \quad (9-9)$$

How to estimate the probability $P(w_i | w_1w_2\dots w_{i-1})$ based on a given data set (corpus) has become a key problem of SLM. It is impossible to get enough data to estimate the $P(w_i | w_1w_2\dots w_{i-1})$, so the n -gram model comes into being, according to which the occurrence of a term is only related to the $n-1$ th term before it (the $n-1$ th term is also called the history of the n th term), that is,

$$P(w_i | w_1w_2\dots w_{i-1}) \approx P(w_i | w_{i-n+1}\dots w_{i-1}) \quad (9-10)$$

When $n=1$, it is referred to as an unigram model, where the occurrence of any term is considered to be independent of other terms, i.e., it is assumed that all terms are independent of each other. A model without regard to term sequence is also called a bag of words model (BOW model). When $n=2$, it is referred to as a bigram model, where the occurrence of the current term is considered to be only related to the previous term.

The basic idea of SLM-based information retrieval models is taking relevance as the sampling probability in statistical models. The earliest model of this kind is the query likelihood model (QLM) put forward by Jay Ponte and Bruce Croft^[13], the main idea of

which is: there's a language model M_d for each document d in the document set, while the query is a sampling result of the model, and documents can be ranked based on the sampling probability of the query based on different document models. The basic formula of the query likelihood model is as follows:

$$\begin{aligned} \text{RSV}(d, q) &= \log P(d | q) = \log \frac{P(q | d)P(d)}{P(q)} \\ &\propto \log(P(q | d)P(d)) \propto \log P(q | d) \\ &= \sum_{t \in q} \log P(t | M_d) = \sum_{t \in q} \text{TF}_{t,q} \cdot \log P(t | M_d) \end{aligned} \quad (9-11)$$

In the above derivation process, the QLM assumes that the prior probability $P(d)$ of the document follows a uniform distribution, i.e., $P(d)=1/N$ (N is the number of documents), so this part can be removed. It should be noted that in some work, the prior probability may be reserved because other distributional hypotheses are adopted. $P(t|M_d)$ is the probability of taking the term t by the model M_d of the document d during sampling. This probability can be calculated by adopting the maximum likelihood estimation (MLE).

$$P_{\text{ml}}(t | M_d) = \frac{\text{TF}_{t,d}}{\sum_{t'} \text{TF}_{t',d}} \quad (9-12)$$

The above estimation is likely to result in zero probability, i.e., the probability that the term t does not appear in the document is estimated to be 0. In this case, once a term in a query does not appear in the document, the score of the document will be zero. This is obviously inappropriate. Therefore, smoothing methods are commonly used for correction. At present, the main smoothing methods include Jelinek-Mercer (JM) smoothing method and Dirichlet Prior smoothing method. The calculation formula of the JM smoothing method is as follows:

$$P(t | M_d) = \lambda P_{\text{ml}}(t | M_d) + (1 - \lambda) P_{\text{ml}}(t | C) \quad (9-13)$$

Where, $P_{\text{ml}}(t|C)$ is the MLE value of t in the entire document collection C . λ is the linear weighting coefficient between 0 and 1, which should be given beforehand. As a result, for every term t in the document collection, it is necessary to calculate the $P_{\text{ml}}(t|C)$. For a document d , the retrieval status value $\text{RSV}(d, q)$ can be obtained by firstly calculating the $P_{\text{ml}}(t|M_d)$ of every term t in the document and secondly getting the $P(t|M_d)$ of every term t in the query q by linear combination. The process is easy to understand, but because the length of this chapter is limited, the concrete calculation example is not presented here. Readers interested in it may have a try by themselves.

A series of other statistical modeling based IR models have been developed based on the query likelihood model, including the KL divergence model^[14], the translation model^[15], etc. With regard to the KL distance model, the KL divergence (relative entropy) between the two kinds of distributions respectively in the query language model and the document language model is calculated, so as to rank documents. The translation model introduces the translation probability between terms, and makes the conversion between terms in the query and those in the document possible.

The SLM-based retrieval model can be derived based on the probability statistical theory. It is highly interpretable and is the most popular retrieval model in the research field. In addition, the model, in essence, does not rely on the term independence assumption, so it is highly extendable. The shortcoming of the model also lies in parameter estimation.

As mentioned before, the query entered by a user might not accurately represent his information need, so it's necessary to expand (or reformulate) the user query in most cases. The most representative method is query expansion based on relevance feedback. The basic idea of the method is using partial documents returned for the first retrieval to modify the original query. The method that the system modifies the query by directly assuming that the top k documents returned for the first time are relevant is referred to as pseudo relevance feedback (PRF). The following part introduces several most classical query expansion methods based on relevance feedback and PRF.

4. Query Expansion Model Based on VSM

The most famous query expansion method in the VSM is the Rocchio method^[16]. The basic idea of the method originates from the assumption that “vector representations of relevant documents are similar to each other, while vector representations of relevant documents are dissimilar to those of irrelevant documents”. Under this assumption, provided that all relevant documents and irrelevant documents in the document collection C composed of N documents are known, and they respectively constitute collection C_r and collection C_{nr} , the optimum query vector that distinguishes the two collections will be:

$$\vec{q}_{\text{opt}} = \frac{1}{|C_r|} \sum_{\forall d_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\forall d_j \in C_{nr}} \vec{d}_j \quad (9-14)$$

Where, $|C_r|$ and $|C_{nr}|$ respectively represent the size of collection C_r and that of collection C_{nr} . The above formula indicates that when all relevant and irrelevant documents are known, the optimum query vector that distinguishes them is the vector difference between the average vector of all relevant documents (centroid vector) and that of all

irrelevant documents^[1].

However, in practice, for a given query, the collection of relevant documents and that of irrelevant documents are unknown beforehand. Although relevance feedback is performed, the relevances of only part of the documents can be obtained. The Rocchio method is a method of gradually modifying the original query vector in the circumstance that the relevances of part of the documents are known. Concretely, assume that \vec{q} is the original query vector, and D_r and D_{nr} are respectively the collection of relevant documents and the collection of irrelevant documents in the returned documents, obtained through relevance feedback or PRF, the query vector after modification will be:

$$\vec{q}_{\text{opt}} = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall d_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_{nr}|} \sum_{\forall d_j \in D_{nr}} \vec{d}_j \quad (9-15)$$

Where, $|D_r|$ and $|D_{nr}|$ respectively denote the sizes of document collections D_r and D_{nr} ; \vec{d}_j is the vector of document d_j ; obviously, $\frac{1}{|D_r|} \sum_{\forall d_j \in D_r} \vec{d}_j$ and $\frac{1}{|D_{nr}|} \sum_{\forall d_j \in D_{nr}} \vec{d}_j$ respectively denote the average vector of all document vectors in the collection of relevant documents D_r and that in the collection of irrelevant documents D_{nr} ; α , β and γ are constants, and are non-negative real numbers.

Consequently, the meaning of the above formula can be described as follows: The query vector after modification is the linear weighted sum of the initial query vector, the average document vector of the collection of relevant documents, and the average document vector of the collection of irrelevant documents, with the weighting coefficients of them being α , β and $-\gamma$ correspondingly. Essentially, the above formula makes the modified query continuously approach the centroid vector of the collection of relevant documents, but gradually deviate from the centroid vector of the collection of irrelevant documents. Due to the subtraction operation in the above formula, the components of the final result vector may be negative. In this case, the commonly used method is to set the value of the components to be 0, i.e., remove the terms which these components correspond to.

In practical applications, there are many value assignment methods for α , β and γ , and a commonly used method is $\alpha=1$, $\beta=0.75$ and $\gamma=0.15$. When $\gamma>0$, it means that the current Rocchio query expansion method allows negative feedback; when $\gamma=0$ and $\beta>0$, it denotes that the current method only allows positive feedback. In addition to the

above mentioned basic Rocchio formula, there are also other transformed Rocchio formulas.

5. Query Expansion Based on Probabilistic Retrieval Model

In the BIR model introduced before, the most important query-related parameters are the probability p_i of occurrence of a term t_i in relevant documents, and the probability q_i of occurrence of a term in irrelevant documents. The query expansion based on such a model mainly refers to the modification to the two parameters.

It is generally assumed that the top k documents in the results returned for the first retrieval are relevant, and they form the collection V of relevant documents, and except V , all other documents in the document collection C are irrelevant. Assume that the document collection in V that contains the query term t_i is V_i , $N=|C|$, $r_i=|V_i|$, and n_i is the number of documents in C that contain t_i , we will get the following formulas based on the definitions of p_i and q_i .

$$p_i = \frac{r_i}{|V|}, q_i = \frac{n_i - r_i}{N - |V|} \quad (9-16)$$

In practice, it is necessary to smooth the above estimations, and a commonly-used smoothing method is adding $\frac{1}{2}$ to the numbers of documents that both contain and do not contain the term t_i . Thus, we get the following formulas:

$$p_i = \frac{r_i + \frac{1}{2}}{|V| + 1}, q_i = \frac{n_i - r_i + \frac{1}{2}}{N - |V| + 1} \quad (9-17)$$

We can find that the original p_i and q_i in the above formulas do not appear in the query updating formula, which is obviously different from the Rocchio method introduced before. A query expansion method that uses the original p_i is as follows:

$$p_i^{(t+1)} = \frac{r_i + \kappa p_i^{(t)}}{|V| + \kappa} \quad (9-18)$$

Where, $p_i^{(t)}$ and $p_i^{(t+1)}$ respectively denote the former p_i value and the p_i value after updating. Essentially, the former p_i value is introduced as the Bayes prior and is used together with the weight κ . Readers interested in it can refer to chapter 11 of the book cited^[1].

6. SLM-based Retrieval Model

A relevance model (RM)^[17] is a query expansion model that is based on the SLM

theory and the PRF idea. In the original statistical language retrieval model, the information on the query's pseudo relevant document collection is not used. However, research shows that this information is very effective in improving the retrieval performance.

The relevance model expands the query by estimating the generating probability $P(t|R)$ of terms in a given query. The formula for calculating $P(t|R)$ is as follows:

$$P(t|R) \approx \frac{P(t, q_1, q_2, \dots)}{\sum_{t' \in V} P(t', q_1, q_2, \dots)} \quad (9-19)$$

and

$$P(t, q_1, q_2, \dots) = \sum_{d \in C_{\text{prf}}} P(d) P(t|M_d) \prod_i P(q_i|M_d)$$

Where, q_1, q_2, \dots are terms in the query, and C_{prf} is the pseudo relevant document collection.

According to the aforesaid formula, the relevance between each term in a term collection and the query q can be reckoned, so as to, in terms of the value scale, select the term set which is most likely to be applied to the expansion, and get new query by adopting the linear interpolation method in combination with both the original and the extended terms. The conditional probability of the new term $P_{\text{new}}(t|M_q)$ is shown as follows:

$$P_{\text{new}}(t|M_q) = (1-\lambda)P_{\text{origin}}(t|M_q) + \lambda P(t|R) \quad (9-20)$$

The relevance model-based sorting function may adopt the KL divergence model^[14] to calculate the KL divergence between $P(t|M_d)$ and $P(t|M_q)$, and to get the final sorting result.

As mentioned earlier, documents and user queries are the basis for the input of the query model, so in selecting the appropriate retrieval model, it is imperative to consider the documents used for retrieval and the characteristics of users' query input. However, the SNS content search is different from traditional text search in both document and query. As a result, corresponding modifications have to be made to query representation, document representation, and relevance computing models in the process of searching. In the following part, we will specifically introduce the SNS content search from the above three aspects.

9.2.2 Query Representation in Microblog Search

Query representation refers to the process of handling queries. In traditional methods, the longer the text, the more sufficient the data for estimating the model, and the more accurate the estimation will be. Generally speaking, if the query and document models are estimated more accurately, the calculation of the relevance will be more accurate, so will the retrieval effect. In the application scenario of microblog search, the microblog documents and queries are both very short, so query expansion becomes more important. The research on query representation in microblog search focuses on query expansion.

In addition, since the real relevance feedback needs abundant user labeling information which is not easy to obtain in practice, the feedbacks are generally the PRFs. The short length of microblog documents leads to the following case: during the implementation of PRF, because the relevant microblog documents are short, when keywords are extracted according to methods similar to TFIDF, the probability that some occasional words are chosen is high; hence, the returned keywords will be noisy, thus greatly affecting the feedback model and the final retrieval effect. In such a case, how to propose a better query representation method and how to use the feedback documents become more important.

1. Query Expansion Based on Internal & External Resources

According to different sources of expansion, query expansions can be divided into the expansions based on external resources and those based on internal resources. The latter refers to the appropriate modification of the original query by using the factors conducive to the performance of microblog retrieval (like URL link, Hashtag, the author, etc.) based on the features of microblog itself after the first retrieval, and the performance of the second retrieval for the new query after the modification; the former generally refers to the expansion of the original query by using the query-related dictionaries, news, Wikipedia and other knowledge bases or results from search engines after the first retrieval. For example, a generative model is used in reference [18] and the external resources are put in the model as variables, where the interdependency among queries, documents and expanded documents is also considered. Specifically, two external resources including news and Wikipedia were used to carry out the experiment in this work, and good results were achieved.

In general, query expansion based on external resources is a commonly-used technology in ordinary searches, and which appears basically the same when being introduced to microblog searches, so we will not go into details about it here. The following part mainly introduces the query expansion based on internal resources. Microblogs involve abundant information, such as the author, Hashtag, URL, etc., which can all be deemed as the resources of query expansion.

Example 9-6 Expansion of microblog queries using Hashtag. In reference [19], the Hashtags in microblogs is mainly used to expand the microblog queries. Specifically, first, in the paper, all Hashtags in the microblog corpus are extracted and the collection of Hashtags is obtained. Then, microblogs containing one certain Hashtag is used to construct the unigram language model of this Hashtag; through calculating the KL divergence between this language model and the query language model M_q , the best k Hashtags are selected for feedback. Finally, the original query model is expanded through the interpolation of the results obtained from the feedback. The final experiment shows that this method has made improvements in metrics like the MAP and $P@10$.

1) Constructing Hashtag model

Assuming that C is a data collection of n microblog posts and in this data collection there are m different Hashtags, denoted as t_i respectively, we collected the microblogs bearing the same Hashtag and estimate the probability of occurrence of a word in each group of microblog posts. Actually, we also achieved m language models, and each language model corresponds to one Hashtag, denoted as Θ_i .

Then, with a given query q (denoted as Θ_q in the model), k Hashtags most relevant to this query is found; the specific approach is sorting each tag t_i based on the following equation:

$$r(t_i, q) = -KL(\Theta_q \parallel \Theta_i) \quad (9-21)$$

After the completion of sorting, the top k Hashtags were selected.

2) Query expansion using Hashtag

It is assumed that r_k is the collection of the top k Hashtags selected as above, and Θ_r is used to represent the corresponding language model, the formula of the post-feedback query model is as below:

$$\hat{\Theta}_{fb} = (1 - \lambda)\hat{\Theta}_q^{ML} + \lambda\hat{\Theta}_r \quad (9-22)$$

Where, $\hat{\Theta}_r$ can be estimated by the following two methods.

- (1) HFB1: Nonzero elements in $\hat{\Theta}_r$ are estimated based on uniform distribution.
- (2) HFB2: Nonzero elements in $\hat{\Theta}_r$ are estimated based on the proportion of $\frac{IDF(t_i)}{\max IDF}$.
- 3) Improvements made using the relationship between Hashtags

Based on the aforementioned expansion models, the relationship between Hashtags will be further considered during feedback. Define X as the matrix of $k \times k$, with each element x_{ij} in the matrix representing the total co-occurrence number of Hashtags t_i and t_j , and carry out normalization for the matrix based on $x_{ii}=1$. Thus, the relevance between the Hashtag t_i and the Hashtag collection r_k can be measured as below:

$$a(t_i, r_k) = \sum_j^k x_{ij} \quad (9-23)$$

If so, when a Hashtag co-occurs with more tags in the r_k , the relevance will be bigger. Introducing the relevance into the above mentioned ranking formula, we get:

$$r_a(t_i, q) = r(t_i, q) + \log a(t_i, r_k) \quad (9-24)$$

Carry out feedback (denoted as HFB1, HFB2 and HFB2a respectively depending on different feedback methods) once more by using the new Hashtag collection calculated in this way. Table 9-2 gives the result of the final query feedback method.

Table 9-2 Experimental result^[19] on microblog retrieval relevance feedback using Hashtag

Method	MAP	NDCG	P10
Baseline, no FB	0.4268	0.6110	0.7034
Baseline, FB	0.4381	0.6209	0.7138
HFB1	0.4605	0.6431 [†]	0.7483
HFB2	0.4617 [†]	0.6388 [†]	0.7414
HFB2a	0.4684 [†]	0.6488 [†]	0.7483

From Table 9-2, we can see that Hashtag can provide useful information for query relevance feedback, and feedback with the consideration of the relationship between Hashtags can further improve the retrieval performance.

Other applications in which internal resources are used: According to reference [20], microblogs containing hyperlinks have more abundant information; therefore, increasing the weight of such microblogs during the PRF will improve the retrieval result.

2. Query Expansion Based on Time Factor

The original query is the direct expression of users' query intent, in which users often use express their query intent with several query words. However, the expected results of queries from Web search and those from microblog search are different. In reference [21], a detailed analysis of queries in microblogs is carried out, and the conclusion is: main purpose of users' searching in microblogs is to query some time sensitive news and social relationship information. In which, the time-based news queries are intended to find the contents that users are interested in from the latest hot microblogs, and these queries are quite sensitive to time.

One typical manifestation of time-sensitive queries is that the distribution of the relevant collections changes significantly at every moment. In reference [22], the queries from No. 301, 156 and 165 TREC are taken as examples for detailed descriptions. In addition, in reference [23], a similar analysis of microblog queries is carried out, to verify the similar characteristics of time response. Figure 9-1 shows the distributions of relevant documents of some randomly selected queries.

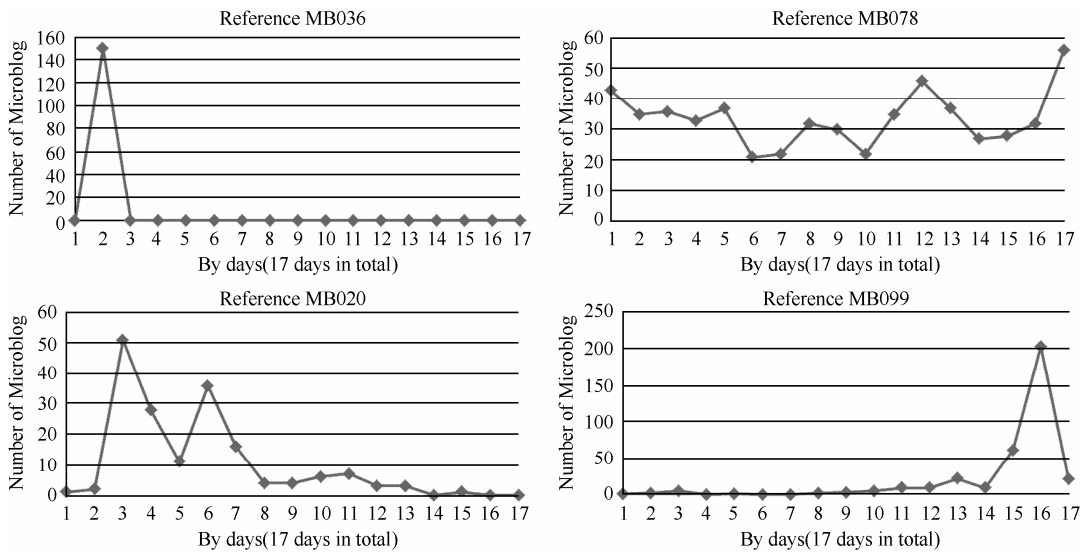


Figure 9-1 Distribution^[23] of Relevant Documents of TREC Queries MB036, MB078, MB020 and MB099 over Time

From Figure 9-1, we can see that the distribution of relevant documents of microblog queries at each moment is far from even and shows obvious time sensitivity. During the

expansion of time-sensitive queries, the time factor should be considered; currently, some time-sensitive query expansion works has sprung up. The works can be divided into two categories based on different ways of time integration: one is to analyze the characteristics of the distribution of documents at certain time points, to select the appropriate time point as the basis of document selection, before carrying out query expansion; the other is to construct the graph by using the time relationship between terms, and obtain the relevance score of each candidate word corresponding to the given query through iterative computations, and then select the terms with high scores for expansion.

Example 9-7 Time-aware query expansion: In reference [24], it is first pointed out that there is correlation between relevance and time to some degree. The author verified this hypothesis by comparing the two kinds of time sequences (one is the time sequence formed by feedback documents consisting of the first retrieval results and the other is the time sequence formed by the really relevant documents). Then, the author counted the number of documents which appeared at each time frame t in the returned document collection at the first time, and then selected the top k documents near the time frame when the largest number of documents appeared, and finally used the Rocchio method to calculate the score of each term, so as to select the expanded word. The experiment proves that this method can improve the retrieval effect.

1) Verifying the correlation between relevance and time

To construct the time sequence $Y=y_1, y_2, \dots, y_n$, the author collected all documents posted in the same time frame t , and the computing method of y_t is as below:

$$y_t = \sum_{d \in D} f_t(d) \quad (9-25)$$

Where, D is the document collection for constructing the sequence and $f_t(d)$ is calculated as below:

$$f_t(d) = \begin{cases} 1, & \text{if } d \text{ is posted in the time frame } t \\ 0, & \text{others} \end{cases}$$

The time frame can be one hour, one day or one month, but in the experiment, the time unit is one day.

By using the above method, the author constructed two time sequences; sequence X stands for the time sequence of real relevant documents while sequence Y represents the time sequence formed by the retrieval results. To determine whether the two sequences are related, the author introduced the Cross Correlation Function (CCF) for calculation:

$$\rho_{xy}(\tau) = \frac{E[(x_t - \mu_x)(y_{t+\tau} - \mu_y)]}{\sigma_x \sigma_y} \quad (9-26)$$

Where, μ_x and μ_y are the mean values of two time sequences respectively, while σ_x and σ_y are the corresponding standard deviations. The τ represents the delay time, $0 \leq \tau \leq 15$. Table 9-3 gives the result of CCF values of the two sequences.

Table 9-3 Result of CCF values of two sequences^[24]

	Full Time				Crawling Time			
	Min	Max	μ	σ^2	Min	Max	μ	σ^2
BL1	0.1216	0.9916	0.7459	0.0364	0.0635	0.9889	0.6122	0.0745
BL2	0.1688	0.9960	0.7145	0.0343	0.0087	0.9957	0.5363	0.0854
BL3	0.1313	0.9943	0.7204	0.0412	0.0586	0.9942	0.5805	0.0832
BL4	0.1577	0.9923	0.7496	0.0368	0.0471	0.9916	0.6043	0.0769
BL5	0.0914	0.9881	0.7094	0.0480	0.1461	0.9917	0.5935	0.0824
	0.1342	0.9925	0.7280	0.0393	0.0648	0.9924	0.5854	0.0805

From Table 9-3, it can be seen that the mean value of CCF is 0.7280 and its variance is 0.0393. Therefore, the retrieved documents and the real relevant documents are highly related in terms of time distribution.

2) Time-based query feedback

According to the above conclusion, the burst time frame of real relevant documents and their initial search results are highly related. Therefore, the author assumed that the pseudo relevant documents at the burst peak period are more likely to be relevant to queries.

In this way, the authors selected the documents at the peak period as the pseudo relevant documents and those at other time periods as the irrelevant documents, and carried out the feedback on queries by using the Rocchio algorithm, with its formula being as below:

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j \quad (9-27)$$

The experiment shows that after the queries are expanded by this method, the retrieval performance can be effectively improved.

Apart from the above work, there's also other work regarding query expansion by

virtue of the time factor.

In reference [24], according to the time T of each document d and the specified rules, the burst collection B (one or more) can be obtained from the first search result collection R of queries, and the collection of bursts is denoted as $\text{bursts}(R)$; then, the score of each term can be calculated according to the following formula:

$$P(t|q) = \frac{1}{|\text{bursts}(R)|} \sum_{B \in \text{bursts}(R)} \left(\frac{1}{N_B} \sum_{d \in B} P(t|d) e^{\gamma \left(\left\lfloor \max_{d' \in B} \text{Time}(d') - \text{Time}(d) \right\rfloor \right)} \right) \quad (9-28)$$

Where, $\text{Time}(d)$ represents the document time and N_B the size of bursts(R) B . At last, the top k terms with the largest probability are selected as the expanded query terms. The experimental result shows that this method can improve the retrieval effect.

In reference [25], in the context of blog retrieval, and based on the idea of the relevance model, a time-based $P(t|q)$ generation model is defined, and the expanded terms ranking ahead are selected through calculation. According to the paper, the generation process of $P(t|q)$ is: first, select a time moment T for the query q and select a t under the conditions of T and q , and thus the formula of the query term generation model is:

$$P(t|q) = \sum_T P(t|T, q) P(T|q) \quad (9-29)$$

The experiment performed on the TREC data collection Blog08 indicates that this method improves the retrieval effect. Similar methods can also be applied to microblog search apparently.

In reference [26], in the background of microblog retrieval, a PageRank-based query expansion method to calculate term weights is proposed. First, extract n -gram from the first returned results as the terms. Then, construct a directed graph for these terms, in which, the terms are taken as the nodes and term TF is the prior value of the node. The edge represents the time correlation between terms, while the weight the time-based relevance. Finally, use the PageRank random walk to calculate the final value of each term, and select some terms to expand the original query. The TREC Twitter data collection-based experiment indicates that good results will be achieved when $n=1$ and $n=2$ are selected for the n -gram.

9.2.3 Document Representation in Microblog Search

The research on microblog document models mainly focuses on the sparsity of

microblogs. On the one hand, the short text is expanded through internal or external resources, to make up for the inadequacy of information, and the effect of microblog retrieval is improved through document expansion; on the other hand, in terms of document representation, the representation method or model suitable for microblog texts is proposed.

1. Document Expansion Based on Internal and External Resources

Many researches have been conducted in the expansion of microblog documents, including expansion based on internal and external resources. For expansion based on external resources, Wikipedia, search engines and other available external data are mainly used; for expansion based on internal resources, documents similar to the original one are retrieved for the purpose of expansion in the context of current data sets, and based on the features of microblogs. Strictly speaking, some work mentioned in this section is not the research directly intended for microblogs. However, during microblog retrieval, the aforesaid method can be used to expand the microblog documents, and then a reranking of those expanded documents is carried out, which might finally improve the retrieval performance. The method based on external resources is conceptually quite similar: namely, expanding the original document based on the external documents, which will not be repeated here. Next, we will mainly introduce the research on the expansion of documents based on internal resources.

Example 9-8 Expansion of documents based on internal resources. According to reference [27], a short text only covers one topic in general. Therefore, in this paper, a short text is regarded as a pseudo query, which is searched in some data collections, and the search results are taken as the object of document expansion.

1) Constructing pseudo queries

First, assume document D as a pseudo query which is denoted as Q_D . Execute query Q_D in collection C to get a ranking result of the top k documents: $R_D=D_1, \dots, D_k$, and their retrieval scores, and then calculate these scores $P(D|D_1), \dots, P(D|D_k)$ by using the QLM.

2) Calculating the expansion document model D'

In a classical statistical language model, the distribution of documents and document collections in vocabularies (in this research, document collections are taken as the background, to play a role of smoothing) is usually used to estimate the language model of documents. If the above ranking result R_D is given, the following formula can be used to improve the estimation:

$$P(w|D') = P(w|d_1, \dots, d_{|D|}) = \frac{P(w, d_1, \dots, d_{|D|})}{P(d_1, \dots, d_{|D|})} \quad (9-30)$$

The denominator $P(d_1, \dots, d_{|D|})$ does not depend on w , so only the joint distribution needs to be retained; namely,

$$P(w, d_1, \dots, d_{|D|}) = \sum_{D_i \in C} P(D_i) P(w, d_1, \dots, d_{|D|} | D_i) \quad (9-31)$$

In which,

$$P(w, d_1, \dots, d_{|D|} | D_i) = P(w | D_i) \prod_{j=1}^{|D|} P(d_j | D_i)$$

According to the above formula, we can get

$$P(w, d_1, \dots, d_{|D|}) = \sum_{D_i \in C} P(D_i) P(w | D_i) \prod_{j=1}^{|D|} P(d_j | D_i) \quad (9-32)$$

The last factor in the aforesaid formula indicates that this joint distribution is, to a large degree, determined by the documents that are most similar to D . Therefore, the k most similar documents can be used to estimate the $P(w | D')$.

After the $P(w | D')$ is gained, the document model can be updated through interpolation:

$$P_\lambda(w | D') = (1 - \lambda) P_{ml}(w | D) + \lambda P(w | D') \quad (9-33)$$

In this paper, $\lambda=0.5$. It should be noted that the time factor is also used in this paper to expand the documents at the same time.

Furthermore, there's also some other works related to expanding the short-text documents. In reference [28], the webpages which the URLs contained in microblogs have links to be used to expand the current microblogs. In some work, the translation model which is expanded based on the integration of microblog Hashtags, URLs and other features is put forward, in which various factors can be conveniently considered, so as to expand the microblog document. The experimental result verifies that microblog documents can be effectively represented by these features.

In addition, it is also possible to realize microblog expansion by using information about authors. In reference [29], an author-based modeling microblog retrieval method is put forward. In the paper, information on authors is extracted from the corpus of the TREC microblog first, and all microblog documents posted by each author are sorted out to form "new documents", and then user's language models are constructed based on these documents; in addition, smoothing is carried out by using the information on authors, to

estimate the probability of the new microblog document terms. The experimental result shows that properly using the author related information can improve the microblog retrieval performance. In reference [30], the followees of the author of microblog documents as well as the microblog documents issued by the followees are used to expand the documents, which also achieved good effects.

2. Document Representation Based on Microblog Features

Example 9-9 Document representation based on microblog features. In reference [31], during the construction of the language model for microblogs, the TF impacts are ignored; namely, the occurrence of terms is denoted as 1 and non-occurrence as 0, while the frequency of terms occurring in a document is not considered. In the paper, the JM smoothing method is used to carry out smoothing for the documents: $P(t|\theta_d) = (1-\lambda)P(t|d) + \lambda P(t)$, where the $P(t|d)$ and $P(t)$ are calculated as follows respectively:

$$P(t|d) = \frac{\hat{n}(t,d)}{\sum_{t' \in d} \hat{n}(t',d)}, \quad P(t) = \frac{\sum_d \hat{n}(t,d)}{N}, \quad \hat{n}(t,d) = \begin{cases} 0, & n(t,d) = 0 \\ 1, & n(t,d) > 0 \end{cases} \quad (9-34)$$

The $n(t,d)$ is the term frequency of term t in document d , and N is the total amount of microblog documents in the data collection.

In the context of microblog search, reference [32] verifies the impacts of such factors as the TF which plays a promoting role in traditional retrieval and Document Length Normalization (DLN) on microblog retrieval. The ranking model adopted in the paper is BM25 model, and the result indicates that the effect is only 0.0048 lower than the best-result $P@30$ when the document's TF is ignored, while the effect will reach the best when the document's DLN factor is ignored. In the end, the paper points out that both TF and DLN are the factors that are involved in the construction of the language model; therefore, how to reduce the negative impacts of these two factors shall be considered during the construction of the language model.

Also, the impact of the time factor can be considered in representing microblog documents. Relevant studies were mainly conducted in the context of statistical language retrieval models, where the key work is to calculate the term generation probability $P(t|M_d)$. There are currently two ways of adding the time factor in the $P(t|M_d)$: one is to define the time weight for the term t by introducing the time factor^[33, 34]; the other is to introduce time during smoothing of the language model probability estimation^[35].

Example 9-10 Temporal language model. In the research work in reference [34], the objective is to determine the document time through the Temporal Language Model (TLM). Although it is not the research about microblog retrieval, the idea regarding introducing time for the term t can be used in the language model. It shall be noted that the data collection at this time will be divided into multiple blocks based on the time granularity, denoted as p , whose set will constitute P . The time granularity can be selected freely, such as 1 month, 3 months, etc. The author defined a time weight in allusion to the term t in the paper, called the Temporal Entropy (TE), and its calculation formula is as below:

$$TE(t_i) = 1 + \frac{1}{\log N_p} \sum_{p \in P} P(p | t_i) \times \log P(p | t_i) \quad (9-35)$$

Where, $P(p | t_i) = \frac{tf(t_i, p)}{\sum_{k=1}^{N_p} tf(t_i, p_k)}$; N_p is the total number of data collection blocks, and

$tf(t_i, p)$ is the number of t_i occurring in the block p . The time weights of terms can be used to modify the term probability of the language model, thus realizing the representation of microblog documents.

In reference [35], the improvement of the document language model lies in that of the smoothing part; the paper holds that the value of smoothing weights varies with the time of different documents; the newer the document, the smaller the λ value; that is, the priority shall be given to the document itself. According to this hypothesis, a new smoothing parameter λ_t is introduced in the smoothing formula and two calculation formulas are given, respectively

$$\lambda_t^{MLE} = \frac{n(d, t < t_d)}{n(C)}, \quad \lambda_t^{MAP} = \frac{n(d, t < t_d) + \alpha - 1}{n(C) + \beta - \alpha - 2} \quad (9-36)$$

Where, $n(d, t < t_d)$ is the number of documents for which the time is smaller than that for document d in the whole data collection; $n(C)$ is the number of documents in the whole data collection; α and β are the conjugate prior parameters of beta.

In addition, there are also many studies in allusion to short-text modeling: some studies[36~39] focus on the classification & clustering and abstract of microblog documents, while some concentrate on the recording training topic model of microblogs, where LDA[40] is adopted. In reference [40], the basic LDA and Author Topic Model as well as the post-expansion LDA are compared, and the comparison result shows that the post-clustering retraining topic model has better effect. All these work can be used to

reevaluate $P(t|d)$, to realize more sufficient representation of documents.

9.2.4 Microblog Retrieval Models

In the context of microblog search, the features of microblogs bring challenges to the retrieval models: first, the text is short; since the calculations in many models are carried out based on statistics, which cannot guarantee the sound effect in the case of super-sparse data. Secondly, the structures are richer; in the original retrieval models, the impacts of repost and reply (the network of microblog documents themselves) and the author's SNS on ranking were not considered. These two problems exist in the vector space model, the BM25 retrieval model and the current language models. Therefore, some changes in the original retrieval models should be made to guarantee the retrieval quality.

1. Time Prior Method

The document prior method is a document processing method where different calculation formulas are defined by taking into consideration the corpus background due to different significances of documents in the corpus, and the calculation results in the retrieval model are used to get the better retrieval effect. At present, researches on the calculation of the document prior in which time information is considered can be divided into two kinds: one is defining changes in the relationship between the document and time; the other is modifying the PageRank method, by adding the time relationship in it. Please be noted that the estimation of document prior here refers to the estimation of $P(d)$, while the document model estimation in the previous section is the estimation of $P(t|d)$.

According to reference [22], prior information varies with the document time. From the perspective of time, the author assumes that the significance of new documents is higher than the old ones, and defines the document significance as the temporal exponential distribution. The formula is as below:

$$P(d)=P(d | T_d) = \gamma e^{-\gamma(T_c-T_d)} \quad (9-37)$$

Where, T_d stands for the time of the document; T_c the latest time for document centralization; γ the exponential distribution parameter, designated according to the human experience. This paper takes the statistical language model as the basis, and regarding the document prior $P(d)$, the original constant is replaced by the exponential distribution (namely, the above calculation formula), and thus the original language model

ranking function is modified. This paper performs the experimental verification on the TREC's news corpus, and the result shows that the time-included ranking result is better than the no-time ranking result. This method can be used in microblog retrieval apparently.

In reference [35], the method put forward in reference [22] is improved; this work points out that the significance of each document varies with different query conditions. Based on this hypothesis, the method of estimating the exponential distribution parameters by using the pseudo relevance feedback collection of queries is put forward; the given query information is introduced through changing the γ in the above formula into γ_q . The pseudo relevance feedback collection of the query q is denoted as $C_{\text{prf}} = \{d_1, d_2, \dots, d_k\}$, and $T_q = \{t_1, t_2, \dots, t_k\}$ is used to represent the moment corresponding to each document in the collection C_{prf} ; according to the maximum likelihood estimation, the calculation formula can be obtained:

$$\gamma_q^{\text{ML}} = 1 / \overline{T_q} \quad (9-38)$$

Where, $\overline{T_q}$ stands for the average mean of the collection T_q , with the value being

$$\overline{T_q} = \sum_{i=1}^k t_i / k .$$

In reference [35], in addition to the verification made on two TREC news corpora, a microblog corpus collection is constructed; the experiment is carried out in the microblog environment, which verifies the retrieval result of this algorithm is better than that of the existing algorithms.

Example 9-11 Time-based statistical language model: In reference [33], a time-based New Temporal Language Model (NTLM) used for webpage searching is put forward, and the queries that this model faces include texts and obvious time information; for example, “earthquakes from 2000 to 2010”. One of the major contributions of this paper is the concept of time-based TF (“TTF” for short) it proposed; namely, the time factor is added to the TF calculation formula, to improve the retrieval effect through time information.

1) Time point extraction

It takes two steps to extract time points. One is the pre-processing stage, in which the paragraph segmentation and word segmentation are carried out. First, extract the main body and distribution time of webpages, and then segment each paragraph into sentences and classify words. Finally, remove the included verbs, conjunctions and prepositions, and the rest of the words will be taken as the keywords.

The other is the extraction stage, which is mainly based on the similarity between sentences and retrospection. First, find the direct or hinted time information from the keywords defined in the previous step; if the time information is not found in the content, take the release time as the webpage's time; for each word w , if time $[t_s, t_e]$ is included in w -covered sentences, we will construct a pair $\langle w, [t_s, t_e] \rangle$ of the keyword and time quantum. If there is no information about time quantum, we will find the best-matching reference time by using the retrospection method; that is, we will find the time of the sentence most similar to this one, and the calculation method of the similarity is as below:

$$\text{Sim}(S_1, S_2) = \frac{\sum_{i=1}^n (k_i \times k'_i)}{\sqrt{\sum_{i=1}^n k_i^2} \times \sqrt{\sum_{i=1}^n k'^2_i}} \quad (9-39)$$

Where, S_1 and S_2 stand for sentences to be matched, $K = \langle k_1, \dots, k_n \rangle$ and $K' = \langle k'_1, \dots, k'_n \rangle$ stand for the frequencies of each keyword occurring in S_1 and S_2 respectively.

2) Time -aware term weight

Generally speaking, the distribution of a term in document d can be estimated with the frequency of this term, as follows:

$$P_{ml}(w | M_d) = \frac{\text{tf}_{w,d}}{\text{dl}_d} \quad (9-40)$$

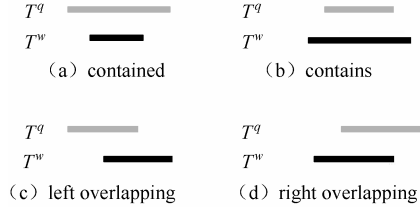
$\text{tf}_{w,d}$ stands for the number of w occurring in document d while dl_d stands for the length of document d . At the same time, the author put forward a term frequency based on the time factor, and its specific definition is shown as below:

$$\text{tf}_{w,d}^T = \frac{\text{num}(w, d, [T_s^q, T_e^q])}{\text{dl}_d} \quad (9-41)$$

$[T_s^q, T_e^q]$ represents the querying time; T_s^q the beginning time and T_e^q the ending time. $\text{num}(w, d, [T_s^q, T_e^q])$ represents that the word w occurs in document d and satisfies $[T_s^w, T_e^w] \in [T_s^q, T_e^q]$. The author defines $[T_s^w, T_e^w] \in [T_s^q, T_e^q]$ as in the four cases as below ($[T_s^w, T_e^w]$ is denoted as T^w and $[T_s^q, T_e^q]$ denoted as T^q).

- (1) Contained: $T_s^q \leq T_s^w$ and $T_e^q \geq T_e^w$.
- (2) Contains: $T_s^q > T_s^w$ and $T_e^q < T_e^w$.
- (3) Left overlapping: $T_s^q < T_s^w$ and $T_e^q \leq T_e^w$.
- (4) Right overlapping: $T_s^q \geq T_s^w$ and $T_e^q > T_e^w$.

The definition of the relationship between the two is as shown in Figure 9-2.


 Figure 9-2 Schematic Diagram of $[T_s^w, T_e^w] \in [T_s^q, T_e^q]$

Finally, in allusion to the above term weight sensitive to time, several smoothing methods are put forward in the paper to improve the retrieval performance.

It is worth noting that TTF can be used in any TF-involved model, such as the language model, BM25, etc., and the experimental result also indicates that TTF can help improve the retrieval effect.

During information retrieval, PageRank (PR for short) indicates the significance of documents by virtue of the links between webpages. References [41~44] are a series of research work where the time factor is added in the original PR algorithm.

In reference [42], it is mentioned that the time dimension information is ignored in the traditional PageRank or HITS algorithm in analyzing the webpage link, based on which, the time factor is added in the original PR algorithm to obtain a new algorithm, called the Timed PageRank (TPR for short), and its formula is as below:

$$\text{PR}^T(T) = (1-d) + d \sum_{T_i \in \text{InLink}T} \frac{w_{T_i} \times \text{PR}^T(T_i)}{C(T_i)} \quad (9-42)$$

Where, $w_{T_i} = b^{(\max \text{time}(C) - \text{time}(T))}$.

In reference [43], the way of adding time is changed and the algorithm T-Rank is put forward. According to Berberich, when a user skips from the current webpage to a link-out page, the skipping probability will be different depending on different link-out webpages; in the meantime, the probability of the user selecting a page at random will vary with the current page, and thereby two probabilities (which can also be called the weight) are defined in the paper: one is the transmission probability of skipping from the page T to a page T_i where the link-out pages of the original page T are centralized, denoted as $t(T_i, T)$; the other is the random skipping weight of the current page, denoted as $s(T)$. The modified formula is as below:

$$\text{TRank}(T) = (1-d) \times s(T) + d \sum_{T_i \in \text{InLink}T} t(T_i, T) \times \text{TRank}(T_i) \quad (9-43)$$

The webpage time is used to define two kinds of freshness which shall be used in the calculation formula of the transmission probability and skipping probability. The experimental result indicates that the effect is improved. In a similar way, in reference [41], the features of the last-modified tab of the webpage are added in the PageRank after being analyzed, to be taken as the weight of the edge between webpages.

In reference [44], the time weight of webpages is added in the score of webpages obtained through the original PageRank algorithm, to adjust the PR. Jing Wan and Sixue Bai believe that a webpage has three properties: integrity, accuracy and time-activity. Time-activity represents the freshness of the webpage content, while outdated information may bring negative impacts during information retrieval, so the author proposed that time-activity shall become one of the factors influencing the PR degree. Assuming a webpage is T , four factors will be used to determine the activity of this page at a certain moment, and these four factors are respectively: the activity of the domain name that this page belongs to, the activity of the creator of the domain that the page belongs to, the degree to which users are interested in this page, and the contents of the text that the page carries. The active scores of this page at different moments are calculated based on these four factors, to form its time activity vector, denoted as $TR(T)$. The author classified the page into four categories: strong-temporal common quality, weak-temporal common quality, weak-temporal high quality and no-temporal quality based on the active score at each moment, and gets the time vector $TR(T)$ of this page; meanwhile, the author gets the PageRank vector $PR(T)$ of this page according to different moments, and the final prior calculation formula of this webpage is $TR(T) \times PR(T)$.

2. Multi-feature Integration Method

In reference [45], microblog search is taken as the research content and the effect of multiple microblog features on microblog search are put forward and verified. The microblog features considered in the paper include: the number of microblogs posted by the microblog author, the number of followers and followees of the microblog author, the microblog length and whether URL is included in the microblog. The numbers of followers and followees of a microblog author are taken as the in-degree and out-degree definition functions to indicate the author's weight. The selected verification method is re-ranking the top k results of special queries in the commercial search engine and determining whether the result after re-ranking is improved. The conclusion obtained finally is that the effect is the best when three features are combined.

Example 9-12 Ranking through integrating multiple features. The idea in reference [45] is simple, with the purpose of re-ranking the initial search results, and the effect of multiple factors on the results is considered.

1) Author score

If A is used to represent the collection of all authors, a mapping $A \rightarrow R^+$ can be established, thus endowing each author a with a non-negative real number $g(a)$, which is called this author's score. First, the score can be estimated through the number of microblogs having been posted by the author so far; the author whose underlying assumption is active may release more meaningful microblogs. Therefore, the author's TweetRank can be calculated through the following formula:

$$TR(a) = N(a) \quad (9-44)$$

The $N(a)$ represents the number of microblogs posted by the author so far. The score can be estimated according to the numbers of followers and followees of author a (FollowerRank), as follows:

$$FR(a) = \frac{i(a)}{i(a) + o(a)} \quad (9-45)$$

The $i(a)$ is the in-degree of the author a and $o(a)$ is the out-degree of the author a . Of course, more complicated algorithms similar to PageRank can also be used to calculate the author's score.

2) Microblog score

Assuming D represents all microblogs and Q represents the set of all queries, the retrieval is equivalent to making such a mapping as $D \times Q \rightarrow R^+$; that is, to get a non-negative real number $f(d, q)$ for a given pair (d, q) . In addition, use "auth" to stand for the author of a document; i.e., $\text{auth}(d)$. Therefore, from the perspective of the author's influence, the microblog score can be estimated through the following two formulas:

$$\begin{cases} f_{TR}(d, q) = TR(\text{auth}(d)) \\ f_{FR}(d, q) = FR(\text{auth}(d)) \end{cases} \quad (9-46)$$

In addition, the length may influence the microblog quality, so LengthRank is calculated as below:

$$f_{LR}(d, q) = \frac{l(t)}{\max_{s \in D_q^k} l(s)} \quad (9-47)$$

The D_q^k is the result of the top k queries returned in the first time, while $l(t)$ and $l(s)$ are

the length of t and s respectively. In addition, URLRank is calculated as below:

$$f_{UR}(d, q) = \begin{cases} c, & \text{if microblog } d \text{ contains URL} \\ 0, & \text{others} \end{cases} \quad (9-48)$$

The c is a positive constant.

3) Combination score

With the combination of the above factors, the FLR (Follower Length Rank) score and FLUR (Follower Length URLRank) score can be obtained:

$$f_{FLR}(d, q) = f_{FR}(d, q) + f_{LR}(d, q) \quad (9-49)$$

$$f_{FLUR}(d, q) = f_{FLR}(d, q) + f_{UR}(d, q) \quad (9-50)$$

The experimental result of the paper indicates that the best method is the FLUR method; that is, to integrate the numbers of the author's followers and followees, the length of the author's microblog, and the factor that whether the microblog contains URLs.

In reference [46], how to use the "learning to rank" method to rank microblogs is introduced, and it is also believed that the microblog search results shall be ranked according to some features rather than the time of the microblogs. To ensure the accurate model, the authors consider many features of microblogs, and then remove some useless ones through feature extraction, principal component analysis (PCA) and other ways, to obtain the conclusion similar to reference [45] which holds that the factor of whether the microblog containing URL, the author's authority and the microblog's length are the factors exerting the most important impacts on the ranking.

In reference [47], it is believed that the difference between queries shall be considered for the "learning to rank" method in the microblog ranking, so they put forward a learning to rank pattern for query modeling; a semi-supervised Transductive Learning algorithm is used to train the model. The learning to rank scheme generally needs labelling. As the Transductive Learning algorithm is used, the ranking of new queries without any labeled examples is also attempted in this paper. The experiment indicates that this method can improve the current "learning to rank" method and improve the ranking effect.

9.3 Content Classification

Classification refers to a process of classifying the given objects into one or more given classes. The text-oriented classification is called Text Classification. Classification

involves two stages: training and test. Simply speaking, training is a process of learning the classification rules or disciplines according to the labeled training data, while test is a process of applying the trained classifiers in new test data. No matter whether it is training or test, characteristic representation shall be carried out for the classification objects first, and then the classification algorithm will be used for learning or classification.

Classification can also be viewed as an application of information retrieval. In this case, the documents to be classified can be taken as “inquiries”, and the target category can be regarded as a “document” in the document set.

Many studies on the short text classification have been carried out at home and abroad. As one of the main features of SNS texts is also “being short”, all the research work about short text classification can be used as reference for the classification of SNS contents. Therefore, the conclusion of this section is not limited to the classification of SNS texts. It should be noted that in the SNS research, according to different classification objects and classification systems, there are also the classification of users, communities, emotions and high/low-quality contents in SNS. Due to the limited space, this section only focuses on topic-based SNS classification. Please refer to Chapter 6 of this book for the emotion classification. Next, the research on classification of current short texts will be summarized from the dimensions of features and algorithms.

Another common content processing technology is clustering, which is also one of the main technologies for topic detection and tracking. Readers who have interest in it can refer to Chapter 11 of this book.

9.3.1 Feature Processing in Short Text Classification

Due to the sparsity of short texts, various resources shall be used to expand the features in classification of texts, so as to expand the short text. In addition, in terms of the feature selection, most researchers will directly adopt the feature selection algorithm in text classification; also, some researchers will combine the features of short texts and modify the current feature selection algorithms or put forward new feature selection algorithms, thus better reducing the short text feature dimension and removing the redundant or irrelevant features, with the hope of obtaining better classification results.

1. Feature Expansion

Most of the microblog text contents are simple and concise, being only a sentence or a few phrases sometimes, and always depending on the context. If the microblog text is too short, the problem of serious data sparsity problem will arise during text classification. To solve the problem, researchers may expand microblogs by introducing external resources; in addition to the WordNet and Wikipedia familiar to us, external resources like the search results, news, Mesh, and Open Directory Project (ODP) obtained through search engines are used as the sources to expand the contents of short texts. It should be noted that the expansion here is also applicable to the microblog document representation mentioned above, except that the final objective of feature expansion is classification.

Example 9-13 Short text classification based on Wikipedia feature expansion method. In reference [48], to overcome the sparsity problem in short text classification, the Wikipedia-based feature expansion method is used to carry out multi-class classification of texts.

(1) Each document d can be represented as the TFIDF form; namely, $\Phi_{\text{TFIDF}}(d) = (\text{TFIDF}_d(t_1), \dots, \text{TFIDF}_d(t_m))$, in which, m is the size of the term space.

(2) Each document maps the short text d into a collection of Wikipedia concepts defined in advance through Explicit Semantic Analysis (ESA). Define the collection of Wikipedia articles as $W = \{a_1, \dots, a_n\}$, to obtain the ESA feature expression of document d : $\Phi_{\text{ESA}}(d) = (\text{as}(d, a_1), \dots, \text{as}(d, a_n))$, in which, the “as” function represents the association strength^[49] between the document d and the concept, and its calculation formula is as below:

$$\text{as}(d, a_i) = \sum_{t \in d} \text{TF}_d(t) \cdot \text{TFIDF}_{a_i}(t) \quad (9-51)$$

(3) Perform classification through two steps: First, search in the whole classification space through calculating the cosine similarity between the document and the class center, to get k candidate classes. Then, use the trained SVM in the k classes to carry out classification.

(4) The above processes can be carried out based on two feature representation methods respectively, and at last, the two result probabilities, P_{TFIDF} and P_{ESA} calculated through the SVM classifier can be integrated, and then the integrated classification probability as below can be obtained:

$$P_{\text{Ensemble}} = P_{\text{TFIDF}} + \alpha P_{\text{ESA}} \quad (9-52)$$

The α is the weighting coefficient, which can be determined in training.

The final experimental result shows that when the Wikipedia features are integrated in

the short texts, the classification effect can be improved significantly.

In reference [50], a method of conceptual expression of short texts based on Wikipedia is also put forward, and the experimental result indicates that this method can improve the classification effect.

Although effective words in short texts have are limited, they contain plenty of latent semantic meanings; how to fully excavate the latent semantic meanings of short texts and the similarity of short texts become one of the major directions of research on short text classification. In reference [51], a latent topics-based short text classification method is put forward; according to this method, latent topics are used to establish a general framework. The latent topics are trained from the “global data set”; Wikipedia and MEDLINE are used in the experiments in this paper, and the test of these two authoritative data sets is carried out respectively. The experimental result shows that this framework can effectively enrich the information in short texts and improve the classification effect.

Under this framework, in reference [52], a method of using the multi-granularity latent topics to enrich short texts is put forward. According to this method, the granularities of topics are divided based on the original method and these generated multi-granularity latent topics are used to enrich short texts, to assist in the classification. At last, both SVM and the maximum entropy model are used for classification, whose error rates of classification are reduced by 20.25% and 16.68% respectively.

There are also many other researches that use external resources to expand short texts; for example, using meta-information in hyperlink targets to improve the classification of microblog topics, as discussed in reference [53]; using Wikipedia to carry out the text feature expansion, as put forward in reference [54]; using other microblogs related to the microblog texts to be classified to enrich the contents of microblog texts for classification..

2. Feature Selection

How to screen out the most representative features is one of the focuses and difficulties of research on short text classification at present.

In reference [56], the method of using eight features of Twitter to classify microblogs into five classes is put forward. These five classes are respectively News, Events, Opinions, Deals and Personal Messages, and the eight features extracted are: Authorship, the presence of abbreviations and slangs, the presence of event related phrases, the presence of emotion words, the presence of emphasizing words (such as *veery*, meaning the high degree of “very”), the presence of symbols of currency or percent, the presence of @ username at the

beginning of the microblog, and the presence of @ username in the middle of the microblog.

The author represented each microblog with the above features, and designed a Naive Bayes classifier to classify the microblogs. The experimental result indicates that the above feature representation has significant improvements relative to the feature representation method for common bag of words, and that the classification has achieved an accuracy of more than 90%.

In reference [57], a simple, scalable and non-parametric method is put forward for the classification of short texts, and its main idea is as follows: first, select some guiding words representative to the topic as the query words, to query the texts to be classified; then, make full use of the information search technology to return the query result; finally, select 5~15 features by voting, to represent and classify the microblogs.

In reference [58], a feature selection method is put forward; first, select some words with rich part of speech as features, and then use Hownet to expand the semantic features of those words, to improve the classification effect. In reference [59], short texts are classified according to features extracted based on the topic trend in Twitter, which achieved certain results.

9.3.2 Short Text Classification Algorithm

1. Improved Traditional Algorithms

In reference [60], it is believed that in the context of microblog classification, classifiers which can realize incremental update are needed. Therefore, the deep analysis of the Bayesian classification method is carried out in this paper and the test on the short text data sets to be classified is carried out; it is found that the classification effect can be greatly boosted by using the effective smoothing method. In addition, the influences of the size and length of the training data collections on the classification effect are also analyzed in this paper.

In addition, the improvements in traditional algorithms are also embodied in the calculation of microblog similarity: the method of calculating microblog similarity is the special similarity calculation method to overcome the problem of huge computation amount during the expansion of microblog contents when external resources are introduced. In reference [61], a “similar kernel function” dedicated to calculating the similarity of short

texts is put forward, and this function can be used in any kernel-based machine learning algorithm; also, the information returned by the search engine can be used to measure the similarity of short texts. In addition, in reference [62], a method of calculating the similarity of two documents that do not share any common terms is put forward, which can be used in microblog classification.

2. Improved Algorithm Based on Concept Drift

Another representative work is to classify microblogs by regarding them as the time-relevant streams. During the time-based microblog classification, the much-discussed problem is the drift of concepts; that is, as time goes on, the topic of classes will vary. Therefore, the original classification model has to be modified constantly.

Some representative work regarding concept drift: in reference [63], the idea of the “stability period” of a word is raised, to overcome the concept drift problem in text classification; in reference [64], the microblog stream is deemed as the research content and a classification model is proposed.

Example 9-14 Classification model of the microblog stream. In this model, in order to deal with the concept drift problem, the global occurrence probability and the recent occurrence probability of each word are estimated and the modeling for common words and some sudden words is carried out; the word changes in its ranking can be detected. Finally, use the word suffix array method to learn the time factor in the n -gram of the word, and present an effective n -gram realization scheme based on the full-text index method. The result of the three data collections-based experiment shows that the change of word probability is a very important factor and that this method can obviously boost the classification effect. In this paper, the Exponentially Weighted Moving Average (EWMA) method is used to estimate the word probability, and its formula is as below:

$$P_{\text{EWMA}}(w_i | c, t) = \sum_{j \in J_c(w_i)} (1 - \lambda)^{|d_{c,t}| - j} \lambda \quad (9-53)$$

The $J_c(w_i)$ is the location collection of a word in $d_{c,t}$ and λ is the smoothing parameter used in EWMA.

Using suffix arrays (SAs) to classify microblogs is a method proposed in recent years. In reference [65], a character string method with good performance in both space and time is put forward based on SAs and the kernel technology in SVM. In reference [66], a new Logistic regression model is presented by virtue of all valid string, and this model is also

constructed based on the SAs technology.

3. Transfer Learning for Short Text Classification

Transfer learning is a framework of machine learning, by which a compact and effective representation can be learned from the labeled data samples in a source domain and unlabeled data samples or a few of labeled samples in the target domain, and then the learned feature representation method is applied in the target domain. The transfer learning technology has been successfully used in many fields, such as the text mining^[67, 68], image classification^[69], named entity recognition^[70], cross-language classification^[71] and WiFi^[72] positioning. In the field of microblog classification, to deal with such features as fast changes in microblog texts and the proneness to outdatedness of training data, quite a few scholars introduced the transfer learning technology, which is briefly summarized as follows.

In reference [73], the Assisted Learning for Partial Observation (ALPOS) is proposed, to solve the short-text problem in microblogs, which is a transfer learning method based on feature representation. According to this method, a microblog text is a part of an ordinary long text, and the reason why the text is short is that some characters have not been observed. In this method, long texts are used as the source data (assisted data), to improve the effect of microblog classification effect. This method expands the framework of the self-learning method, requiring that the source data and the target data have the same feature space and tag space, and the labeled source data are necessary.

Advantages of the transfer learning method lie in its simple operation and significant effect, while the disadvantages are that it is only applicable to situations where the source field and the target field are quite similar, and it also requires that the source field and target field have the same class space and feature space; the efficiency of the iteration-based training process is not high. When some instances in the source field in the application scene conform to the distribution of the target field, the instance-based transfer learning method is applicable.

In reference [74], a transfer learning method is proposed to solve the problem of transfer classification of short & sparse text (TCSST); this method is an instance-based transfer learning method in the inductive transfer learning and expands the TrAdaBoost framework. To solve the problem of sparse labeled data in the source data, the semi-supervised learning method is adopted to carry out sampling for the original data. Based on the original labeled data, post-sampling original unlabeled data and target labeled

data, and the TrAdaBoost's framework training classifier, the method performed experimental verification on the 20-Newsgroups data collection and a real seminar review datum.

9.4 Social Network Recommendation

Recommendation is a process of pushing relevant items or contents to users according to their preferences. It mainly includes the content-based recommendation technology and the collaborative filtering-based recommendation technology. The basic idea of content-based recommendation technology is to extract the characteristics of users from the personal information of users or to analyze the preferences of users to items according to the historical data of users, then build a preference model of users by synthesizing these factors, and finally calculate the degree of matching between the items to be recommended and users based on these characteristics. Another practice is to look for other users with similar preference or similar items from the score data of large number of users, and then use such similarities to recommend items for current users; this practice is referred to as collaborative filtering.

Recommendation is a common application form of information retrieval. In this application, a user can be taken as an “inquiry”, an item as a “document”; therefore, recommendation is a process to obtain a “document” best matching the “inquiry” from these “documents”. Thus, the basic model and algorithms of information retrieval can both be applied in a recommendation system, especially a content-based recommendation system. A recommendation system based on collaborative filtering is somewhat different from an ordinary information retrieval system. Therefore, in the presentation below, we will mainly present the recommendation system based on collaborative filtering.

Generally speaking, a recommendation system contains the following three elements.

User: the target of a recommendation system. The recommendation effect can vary greatly with the amount of user information mastered by the system. User information includes a user's personal attributes, such as age, gender and occupation, as well as historical exchange data of the user with the system, which can more directly reflect the user preference.

Item: the contents of a recommendation system. Here, item is a concept in a broad sense, it can be both commodities such as books, music and films, and also information contents such as news, articles and microblog.

Rating of user preference for items: in websites like Amazon or Netflix, scores are normally used to indicate user's preference for items. The commonly used scores are divided into five levels (1~5), with ratings from low to high. Depending on applications, the rating of preference degree can be in different forms, and sometimes two levels (like or dislike) are used for the rating.

Recommendation algorithm: given user u and item i , estimate the rating r of a user for the item. The recommendation algorithm can be abstracted as function $f(u,i)=r$. Essentially, the recommendation algorithm provides the estimation of matching degree of user and item; to obtain a precision recommendation result, it is necessary to tap in-depth the characteristics of the user and the item, and the interactive relationship between them. The recommendation algorithm studies how to model various potential factors that may affect the recommendation.

As mentioned above, recommendation algorithms can be classified into two major categories according to different information used.

(1) Content-based recommendation: the basic idea is to analyze the personal attributes of users and the attributes of item, calculate the matching degree between the item to be recommended and the user in conjunction with the historical preference of the user for the item, so as to recommend an item with high matching degree to the user. This technique allows directly completing recommendations based on information of users and items, and was used extensively in the early recommendation systems; however, this technique can only push items similar to historical characteristics of items by establishing a direct relationship between item characteristics and user preference, but it cannot produce novel contents, and the expandability is poor[75~80].

(2) Recommendation based on collaborative filtering: the evaluation of users on items constitutes the user-item matrix (see Table 9-4). Based on the basic assumption that "similar users have similar preferences", items preferred by users similar to the target user can be pushed to the latter. Collaborative filtering method does not take into account contents, and is entirely based on the historical information on interactions between users and items; therefore, it can solve the problem of lacking user and item description information in many cases. With its effectiveness in recommendation, collaborative filtering is now extensively applied in various commercial systems, such as, news recommendation in GroupLens, movie recommendation in MovieLens and music recommendation in Ringo.

Table 9-4 User-item matrix

	Item 1	Item 2	Item 3	Item 4
User 1	1		2	
User 2		2		2
User 3	2	3	4	
User 4		2		5

In an actual system, to obtain better recommendation results, the two methods can be combined, such as the MatchBox recommendation system of Microsoft^[81]. In the research circle, collaborative filtering has attracted extensive attention for two reasons: one is that the inputs (user-item matrix) of collaborative filtering is more accessible to many systems, while it is difficult to obtain users' personal attributes (except for a small number of social network websites); the other is that it has higher theoretical values to mine the user-item matrix by integrating multiple sources of information. Due to limit of space, this section is mainly focused on how social network information is integrated in the collaborative filtering technology. In the following part, the meaning of social recommendation will be presented first, and then the memory-based social recommendation and model-based social recommendation method will be presented.

9.4.1 Brief Introduction to Social Recommendation

This section presents the differences between social recommendation and traditional recommendation technologies based on collaborative filtering, so as to understand the basic issues of social recommendation.

1. Problems of Traditional Collaborative Filtering

For the given user-item matrix $\mathbf{R} \in R^{m \times n}$, where m is the number of users, and n the number of items; the score of user u for item i is $r_{u,i}$ (or $R_{u,i}$), corresponding to the element in the u th line and i th row of \mathbf{R} matrix. In the actual applications, most elements in matrix \mathbf{R} are missing, so the target of collaborative filtering is to use the existing elements in \mathbf{R} to predict the missing elements $\tilde{r}_{u,i}$ in \mathbf{R} ; i.e. to predict the rating of user u for item i .

The two main problems faced by collaborative filtering are:

(1) Data scarcity: in the user-item matrix \mathbf{R} , all known elements come from user operation, but in real applications, both the number of users m and number of items n are huge, while items really scored by users are quite limited in number; for example, in the Netflix data, only about 1% of the score values are known.

(2) The problem of cold start: collaborative filtering uses the historical score information of users to make recommendation; if a user uses the recommendation system for the first time, there is no historical information available at that time; therefore, it is quite difficult to recommend items to such users. Solutions to cold start include using the average value of existing user scores in the matrix to predict new users' scores for fore items, or forcing users to score some items to obtain available historical data from the users.

2. Problems With Social Recommendation

The ultimate goal of social recommendation is the same as that of collaborative filtering, i.e. to predict the missing items in the user-item matrix $\mathbf{R} \in R^{m \times n}$; the available information includes two categories:

- (1) User historical score information $\mathbf{R} \in R^{m \times n}$;
- (2) The social relation information of users.

Social relation is defined as a function on the user set; for example, in a social relation network $G = \langle V, E \rangle$, V is a set of users, $\langle u, v \rangle \in E$ indicates a connecting side of user u and user v , the social relation based on a social relation network can be expressed using user -user matrix $\mathbf{T} \in R^{m \times m}$; i.e., if there is a relationship between user u and user v , then

matrix element $T_{u,v} = 1$; otherwise $T_{u,v} = 0$.

Obviously, social network relation is just one of the in the social relations, and social relations extracted based on different applications are also different. For example, in the Epinions website^[82], users can not only evaluate various items, but also score the evaluations made by other users. In addition, the Epinions website uses the concept of “trust”, and users can choose to directly “trust” or “screen” some users; in this way, every user has his or her own trusted target user, and can be also possibly trusted by other users, thus forming a complicated trust network (see Figure 9-3).

Despite their diversified forms, social relations have the potential of improving the

effect of collaborative filtering as long as they can reflect the similarities between users, because the basic assumption of collaborative filtering is to make recommendation by making use of the preference of similar users. Social recommendation makes use of the social relations between users to improve the effect of collaborative filtering.

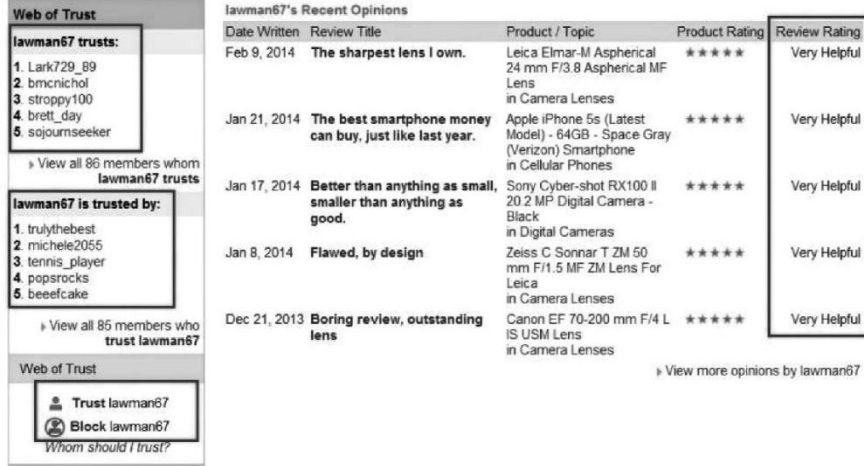


Figure 9-3 Direct trust relations between users in Epinions website

3. Evaluation for Collaborative Filtering

There are two commonly used evaluation metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Both are used to measure the degree of difference between the predicted values and true values. The calculation formulas corresponding to the two metrics are as follows:

$$\begin{cases} \text{MAE} = \frac{1}{T} \sum_{u,i} |\tilde{r}_{u,i} - r_{u,i}| \\ \text{RMSE} = \sqrt{\frac{1}{T} \sum_{u,i} (\tilde{r}_{u,i} - r_{u,i})^2} \end{cases} \quad (9-54)$$

Where, $\tilde{r}_{u,i}$ and $r_{u,i}$ represent respectively the predicted values and true values, and T the number of elements predicted.

9.4.2 Memory Based Social Recommendation

The memory based collaborative filtering algorithm^[40, 83~88] requires complete

user-item matrix in calculation, and there are two basic assumptions:

(1) Similar users (based on historical record of users' scores for items) give similar scores for new items.

(2) Similar items (based on historical record of users' scores obtained) receive similar scores from new users.

Recommendation based on the first assumption is also referred to as user-based collaborative filtering, and that based on the second assumption is referred to as item-based collaborative filtering. Essentially, both assumptions filter off irrelevant score information, and use the most similar user-item score information to make prediction for $r_{u,i}$.

1. User-based Collaborative Filtering

The predicted score of user u for item i can be expressed as the formula below:

$$\tilde{r}_{u,i} = \bar{r}_u + \frac{\sum_{v \in N(u)} \text{sim}(u,v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N(u)} \text{sim}(u,v)} \quad (9-55)$$

Where, $N(u)$ indicates the set of users most similar to user u , and the number of set $|N(u)|$ should be designated in advance; \bar{r}_u is the average value of user u 's scores for all items; and $\text{sim}(u,v)$ gives the similarity between user u and user v . The similarity calculation $\text{sim}(u,v)$ is a critical step in the memory based collaborative filtering algorithm, and commonly used similarity calculation methods are mainly Pearson's correlation coefficient^[40] and cosine similarity^[85].

Pearson's correlation coefficient is used to reflect the degree of linear correlation of two variables, and the calculation formula is as follows:

$$\text{sim}(u,v) = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}} \quad (9-56)$$

Calculation of $\text{sim}(u,v)$ also includes the limited Pearson's correlation coefficient^[89], Spearman's ranking correlation coefficient^[90] and so on. The following is a specific example of calculation.

Example 9-15 User-based collaborative filtering method. Suppose the user-item score matrix is as it is shown in Table 9-5, in which the scores given by users for items are integers of 1 to 5, we will use the user-based collaborative filtering algorithm to predict users' scores for items.

Table 9-5 Example of user-based collaborative filtering

User \ Item	I_1	I_2	I_3	I_4	I_5
U_1	4		5		4
U_2	4	2	5		1
U_3	2	4		2	2
U_4	1	5		3	3

Suppose it is required to predict the score of user U_1 for item I_2 , according to the user-based collaborative algorithm, we can use the above-mentioned score values to calculate the similarity of the other three users U_2, U_3, U_4 to user U_1 :

$$\text{sim}(U_1, U_2) = \frac{(4 - 4.3)(4 - 3) + (5 - 4.3)(5 - 3) + (4 - 4.3)(1 - 3)}{\sqrt[2]{(4 - 4.3)^2 + (5 - 4.3)^2 + (4 - 4.3)^2} \cdot \sqrt[2]{(4 - 3)^2 + (5 - 3)^2 + (1 - 3)^2}} = 0.6923$$

In the same way, we can obtain

$$\text{sim}(U_1, U_3) = 0.5183$$

$$\text{sim}(U_1, U_4) = 0.3$$

Therefore, the score value of user U_1 for item I_2 can be calculated as follows:

$$\begin{aligned} & r_{U_1, I_2} \\ &= \bar{r}_{U_1} + \frac{\text{sim}(U_1, U_2) \cdot (r_{U_2, I_2} - \bar{r}_{U_2}) + \text{sim}(U_1, U_3) \cdot (r_{U_3, I_2} - \bar{r}_{U_3}) + \text{sim}(U_1, U_4) \cdot (r_{U_4, I_2} - \bar{r}_{U_4})}{\text{sim}(U_1, U_2) + \text{sim}(U_1, U_3) + \text{sim}(U_1, U_4)} \\ &= 4.3 + \frac{0.6923 \cdot (2 - 3) + 0.5183 \cdot (4 - 2.5) + 0.3 \cdot (5 - 3)}{0.6923 + 0.5183 + 0.3} = 4.7535 \end{aligned}$$

2. Item-based Collaborative Filtering

Similarly, the similarity between items can also be used to predict the score given by user u to item i , and in this case, the calculation formula is

$$\tilde{r}_{u,i} = \bar{r}_i + \frac{\sum_{j \in N(i)} \text{sim}(i, j)(r_{u,j} - \bar{r}_j)}{\sum_{j \in N(i)} \text{sim}(i, j)} \quad (9-57)$$

In the same way, $N(i)$ indicates the set of items most similar to i , and the number of set $|N(i)|$ should be designated in advance. \bar{r}_j is the average value of scores obtained by

item j from all users. The calculation of item similarity $\text{sim}(i, j)$ is similar to that of user similarity $\text{sim}(u, v)$, and similarity calculation methods such as Pearson's correlation coefficient can also be used.

3. Integrating Social Network Information

User-based collaborative filtering transmits scores by estimating $\text{sim}(u, v)$, and smoothens the scores of current users by using scores of neighbor users. However, in the real environment, as the score data of most users is limited, the problem of data sparsity is quite serious, making it very difficult to make similarity calculation for many users. Typical social network information includes association relationship between users, reflecting to a certain degree the similarity or trust between users. In fact, researches in social science show that trust diffusion exists extensively in social networks, as well as the trust-based socialized recommendation^[91]; therefore, modeling the degree of user trust is an important task for social recommendation.

In references [76] and [83], the user trust relationship in Epinions is used to improve the calculation of user similarity. In references [92] and [93], the factor of trust relationship is introduced into the recommendation system, to adjust the weight of scores with trust degree, and the experimental results have verified that user trust relationship can improve recommendation quality. Reference [90] proposes the trust relationship based semantic web recommendation system, which is based on intellectual entities, and connections are established between entities based on trust relationship.

Example 9-16 Recommendation based on user friend relations: reference [94] proposes a social relation diagram model constructed based on the friend relations of users and social tag information, and proposed a random walk model RWR, to integrate user friend relations and social tags, thereby increasing the effect of recommendation (See Table 9-6).

Table 9-6 Example of user relations

User \ User	U_1	U_2	U_3	U_4
U_1	0	3	0	5
U_2	2	0	7	0
U_3	5	2	0	7
U_4	0	3	6	0

The basic idea of memory-based social recommendation methods is that users and their friends share similar preferences; therefore, the trust degree between users is used directly to substitute the score similarity between users. Take the user trust relationship in Table 9-6 as an example, the user-based collaborative filtering integrated with user relations is calculated as follows:

$$r_{U_1, I_2} = \frac{\text{trust}(U_1, U_2) \cdot r_{U_2, I_2} + \text{trust}(U_1, U_4) \cdot r_{U_4, I_2}}{\text{trust}(U_1, U_2) + \text{trust}(U_1, U_4)} = \frac{3 \times 2 + 5 \times 5}{3 + 5} = 3.875$$

This method relies upon the trust relationship between users and their friends; however, it can be seen from the example above that, there is no direct trust value between user U_1 and user U_3 , and when such relationship is seriously missing, the effect of this algorithm will be affected; therefore, researchers proposed establishing indirect trust relationship for users through trust transfer.

Reference [95] proposes a trust-based recommendation method, to calculate user's authority scores through trust transfer, and use the hierarchy of trust relationship to substitute the similarity calculation in traditional recommendations. Reference [96] proposes a recommendation method based on the random walk model, which can provide the credibility of score results.

In the recommendation methods based on trust relationship, it is most critical to obtain the

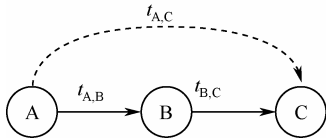


Figure 9-4 Transmissibility of trust relationship

set of trusting users, thereby avoiding the problem of data sparsity in the nearest neighbor based method. It is generally believed that trust between users is transferrable, as shown in Figure 9-4.

If User A trusts B, and B trusts C, then it is believed that A also trusts C. In addition, trust relationship is one-way, if A trusts C, it does not

mean that C also trusts A. Through the transfer of relationship, a user's range of trust relationship can be effectively expanded, even in a case of very sparse data; for example, user A only has one directly related trusting user B, but through layers of transfer by the trusting users of B, a fairly big set of trusting users for user A can still be obtained, to effectively relax the data sparsity problem in score prediction.

A classical practice is to combine the maximum transfer distance with the minimum trust threshold^[75], to limit by maximum transfer distance the levels of trust transfer in the calculable range, and then sieve out with minimum trust threshold the trusting friends that can be used.

Reference [77] proposes a path-algebra-based trust transfer model; similarly, reference [78] proposes a method to express user trust relationship with trust transfer matrix, which can effectively integrate a number of trust transfer models. Reference [97] proposes the TidalTrust, to make recommendations by using the breadth-first search strategy in the trust relationship network. First, the shortest distance from all score users to target users is calculated, and then the weight of scores is adjusted with the trust relationship between users on the shortest distance path. This approach only uses users on the shortest path, and is very likely to lose a wealth of valuable information. On the basis of research work in reference [97], reference [76] puts forth the MoleTrust, by introducing the concept of maximum trust depth; it requires that the integrated trust users are not only limited within the shortest distance, but also within the maximum depth range, equivalent to a compromise between precision and coverage. Reference [98], based on the spreading activation model, puts forth the AppleSeed model, which states that the trust relationship of users has the cumulative effect. Even if the trust path between two users is very weak, it is also possible to obtain a very high trust weight as long as there are sufficient connection paths. In the following part, we will take TidalTrust as an example to introduce the trust transfer process.

Example 9-17 The method to adjust score weight with trust relations (TidalTrust).

Based on the trust relationships between users in Table 9-6, the trust diffusion path from user U_1 to user U_3 can be obtained, as shown in Figure 9-5.

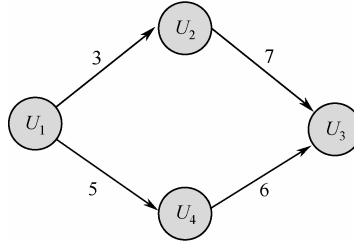


Figure 9-5 Trust diffusion path from user U_1 to user U_3

According to TidalTrust, it is necessary to first calculate the maximum value of trust at the nodes, and take this maximum value as the threshold to select a usable node, the actual calculation method is as follows:

$$S_{ij} = \frac{\sum_{u_k \in \mathcal{F}(i) \geq \max S_{ik}} S_{ik} S_{kj}}{\sum_{u_k \in \mathcal{F}(i) \geq \max S_{ik}} S_{ik}} \quad (9-58)$$

where, $u_k \in \mathcal{F}(i)$ represents all neighboring friends of user i , and some neighboring

friends with trust degree above the threshold are selected by $u_k \in \mathcal{F}(i) \geq \max$ as the nodes of diffusion.

It can be seen from Figure 4-5 that, in this example, U_2 and U_4 are the neighboring nodes of target user U_3 ; the trust value from user U_1 to U_2 is 3; the trust value from U_1 to U_4 is 5; only U_1 is connected to U_2 and U_4 ; therefore, U_2 and U_4 are respectively marked as 3 and 5. As for U_3 , the threshold of trust degree for the node is 5; therefore, only U_4 satisfies the conditions. Thus, we can obtain the trust degree of users U_1 and U_3 as follows:

$$\text{trust}(U_1, U_3) = \frac{5 \times 6}{5} = 6$$

Apply this trust value in the above-mentioned user-based collaborative filtering formula, to achieve the score value of U_1 for item I_2 as follows:

$$\begin{aligned} r_{U_1, I_2} &= \frac{\text{trust}(U_1, U_2) \cdot r_{U_2, I_2} + \text{trust}(U_1, U_3) \cdot r_{U_3, I_2} + \text{trust}(U_1, U_4) \cdot r_{U_4, I_2}}{\text{trust}(U_1, U_2) + \text{trust}(U_1, U_3) + \text{trust}(U_1, U_4)} \\ &= \frac{3 \times 2 + 6 \times 2 + 5 \times 5}{3 + 6 + 5} \approx 3 \end{aligned}$$

All the above-mentioned researches are memory based methods, in which recommendations are made by using user relations based on the heuristic rules, which cannot optimize the effect. To solve this problem, researchers put forth many model-based methods, to obtain parameters of models through machine learning.

9.4.3 Model Based Social Recommendation

The model-based recommendation algorithm models the user-item matrix, to learn from it the corresponding model parameters, so recommendation results can be obtained by just loading the model, without the need to analyze the original user-item matrix. Common recommendation models include the neighborhood model and the matrix decomposition model.

1. Neighborhood Model

The neighborhood model was an improvement on the memory-based recommendation method; it is proposed in reference [99], with the basic idea to obtain similar users on the basis of the score similarity between users, and then predict users' scores for items by using the scores given by similar users. The calculation formula is as follows:

$$\tilde{r}_{u,i} = \sum_{v \in V} r_{v,i} w_{u,v} \quad (9-59)$$

where, $w_{u,v}$ is the interpolation weight, which is used to integrate score values of different users. This idea is basically the same as that of the memory-based collaborative filtering algorithm, except that after selecting similar users, it does not use similarity as the weight to integrate the score values of users; instead, a regression model is used to obtain the values of these weights through learning and training, and the optimized target function is expressed as follows:

$$\min \sum_{v \in V} \left(r_{v,i} - \sum_{j=1}^n r_{v,j} w_{u,v} \right)^2 \quad (9-60)$$

The interpolation weight parameters obtained through learning can better fit the data deviation, and improve the recommendation effect.

Some scholars introduced the bias information on the basis of this job^[100, 101].

$$\tilde{r}_{u,i} = \mu + b_u + b_i + \sum_{j \in R(u)} (r_{u,i} - b_{u,j}) w_{i,j} + \sum_{j \in N(u)} c_{i,j} \quad (9-61)$$

In this model, μ is the global average score, b_u the bias of user, b_i the bias of item, $R(u)$ the set of items already scored by user u ; the parameter w is the tightness between items; $N(u)$ is the set of items for which user u has made implicit feedback; $c_{i,j}$ represents the bias based on implicit feedback item j on item i .

2. Matrix Decomposition Model

The basic idea of this model is to decompose the user-item score matrix. After the decomposition, both users and items can be expressed with k -dimension implicit characteristics. Matrix decomposition can reduce the dimensions of users' score data on items, to build the characteristic models of users and items with less dimensions, thereby effectively solving problems resulted from data sparsity. Many recommendation models based on matrix decomposition are now available^[83, 92, 93, 97, 102].

The RSVD model^[103] is one of the classical matrix decomposition methods. Specifically, the user-item matrix \mathbf{R} can be decomposed into a form of user matrix \mathbf{U} multiplied by item matrix \mathbf{V} , as shown by the following equation:

$$\mathbf{R} = \mathbf{U}^T \mathbf{V}$$

where, $\mathbf{U} \in R^{k \times m}$ and $\mathbf{V} \in R^{k \times n}$ respectively express the k -dimension implicit characteristic vectors of users and items; by minimizing the residual differential square, the expression of

users and items in implicit space is obtained:

$$\mathbf{U}, \mathbf{V} = \arg \min_{P, Q} O(\mathbf{U}, \mathbf{V}) = \arg \min_{P, Q} \frac{1}{2} \|\mathbf{R} - \mathbf{U}^T \mathbf{V}\|_F^2 + \lambda (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \quad (9-62)$$

Where, λ is the coefficient of a regular term, which can prevent overfitting of the model. This formula can be solved by using gradient descent method^[83, 93]. The RSVD model is actually the basic frame of matrix decomposition, which has large room for improvement. To better fit a user-item matrix, reference [104] introduced global bias, user bias and item bias, so the scoring term can be expressed as

$$R_{u,i} = \bar{r} + b_u + b_i + \sum_{k=1}^K \mathbf{U}_{u,k}^T \mathbf{V}_{i,k} \quad (9-63)$$

where, \bar{r} is the global average score, b_u is the user bias, and b_i the item bias. The corresponding optimizing target function is expressed as

$$O(P, Q) = \frac{1}{2} \sum_{u=1}^m \sum_{i=1}^n (\tilde{r}_{u,i} - r_{u,i})^2 + \lambda_1 (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + \lambda_2 (b_u + b_i)^2 \quad (9-64)$$

where, $\|\mathbf{U}\|_F$ and $\|\mathbf{V}\|_F$ are respectively F norms of matrixes \mathbf{U} and \mathbf{V} , i.e. the squares and square roots of all elements of the matrixes. λ_1 and λ_2 are combinatorial coefficients. Paterek has very good universality for expanding ideas of the model. There may be many potential factors influencing the recommendation result, such as, the information about time and geographic locations, which can all be added into formula (9-60) as bias terms, so as to better fit the user-item score value. Reference [105] considered the matrix decomposition from the perspective of probability, and put forth the probability matrix decomposition model PMF, which assumes that user's score R for items conforms to the normal distribution with the mean value $P^T Q$ and variance σ_R^2 , and that a set of users with similar scores have similar preferences.

3. Integrating Social Network Information

Essentially, two assumptions are associated with the influence of social networks on recommendation:

- (1) User scores are subject to the influence of the user's friends.
- (2) The related user implicit vector expressions should be similar.

Recommendation methods integrated with social relations of users can make effective use of user relations, and the currently prevailing methods can be classified into three types.

(1) Common factorization model: methods of this type factorize the user-to-item score matrix and user relation matrix at the same time, and both matrixes share the same user bias vector. The basic assumption is that users have the same bias vector in the score matrix space and relation matrix space, and this method is used to merge the score data with user relation data. Representative models include SoRec and LOCABAL.

(2) Integration models: methods of this type make weighed merging of the friend-to-term score matrix decomposition term with the user score matrix decomposition term. The basic assumption is that users have similar preferences as their friends; therefore, the bias term of users is smoothened with bias term of friends. Representative models include STE and mTrust.

(3) Regular model: methods of this type introduces regular terms of user relations in the matrix decomposition model, and in the optimizing process, the distance between user-and friend's bias terms is also one of the optimizing targets. Its basic assumption is similar to that of integration models, and users have similar bias as their friends. The representative models include SocialMF and Social Regularization.

Example 9-18 The graphic model method SoRec integrating social relations and score records. Reference [99] put forth SoRec, which integrates the social relations and score records of users by mapping the user social relations and user score records on the same implicit characteristic space, thus mitigating the data sparsity and improving precision. The graphic model of SoRec is as shown in Figure 9-6.

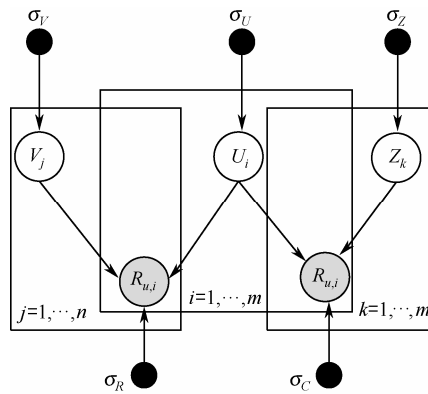


Figure 9-6 Graphic model expression of SoRec

As shown in Figure 9-6, the SoRec model factorizes the user score matrix into two terms U_i and V_j , and the user relation matrix into two terms U_i and Z_k ; U_i as the common

term, links the two matrixes, thus realizing the integration of user relations. Its optimizing target function is as follows:

$$\min \| \mathbf{R} - \mathbf{U}^T \mathbf{V} \|^2 + \alpha \sum_{i=1}^n \sum_{uk \in N_i'} (S_{ik} - u_i^T z_k)^2 + \lambda (\| \mathbf{U} \|^2 + \| \mathbf{V} \|^2 + \| \mathbf{Z} \|^2) \quad (9-65)$$

Based on assumption 1, if two users have a friend relationship, their scores for items should be close to each other. Reference [106] put forth the STE model, which performs linear combination of the basic matrix factorizing model with social networks, and makes final score prediction by weighted summation of basic score matrix and trust score matrix. Its graphic model expression is as shown in Figure 9-7.

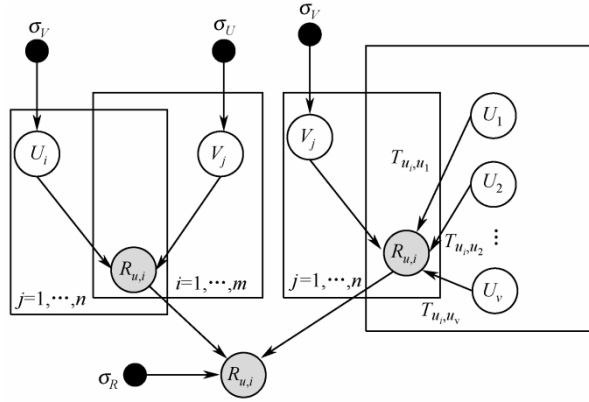


Figure 9-7 Graphic model expression of STE

As shown in Figure 9-7, the STE model first factorizes the user-to-item score matrix, then makes weighted merging of user's bias term on an item and the bias term of friend on the same item, with the weight value being the weight of user's relation with friend.

$$\mathbf{R}_{u,i} = g(\alpha \mathbf{U}_u^T \mathbf{V}_i + (1 - \alpha) \sum_{v \in N_u} T_{u,v} \mathbf{U}_v^T \mathbf{V}_i) \quad (9-66)$$

The corresponding $\mathbf{R}_{u,i}$ is the normalized user scores, and g (g) is a logistic function. Obviously, reference [106] is only a linear combination model of a superficial level, and has not tapped deeply the bonding role of friend relations. For reference [106] and reference [99], the characteristic space of neighbors affects the user's score, rather than its characteristic space; therefore it is not possible to solve the issue of trust transfer.

Based on assumption 2, if the two users have relations in social network, the user vector expressions after matrix decomposition should have some similarities. Assumption 2 can be taken as a restriction on the matrix decomposition, equivalent to adding an extra

regular term into the previous optimizing target, as indicated below:

$$O(U, V) = \frac{1}{2} \|R - U^T V\|_F^2 + \lambda (\|U\|_F^2 + \|V\|_F^2) + \sigma L(U) \quad (9-67)$$

Where, $L(U)$ is the regular term of social networks to user implicit vector. The relevant work can better fit the recommendation result by designing the mathematic form of $L(U)$. Specifically, reference [107] has proposed two forms of regularization as follows:

$$L(U) = \sum_{i=1}^m \left\| U_i - \frac{\sum_{f \in F^+(i)} \text{sim}(i, f) U_f}{\sum_{f \in F^+(i)} \text{sim}(i, f)} \right\|_F^2 \quad (9-68)$$

$$L(U) = \sum_{i=1}^m \sum_{f \in F^+(i)} \text{sim}(i, f) \|U_i - U_f\|_F^2 \quad (9-69)$$

where, $F^+(i)$ is the collection of friends of user i , $\text{sim}(i, f)$ is the similarity between user i and user f . The experiment indicates that regularization based on the second formula produces better effect.

In the PLSF model proposed by Shen^[108], the regular term used is in the form below

$$L(U) = \sum_{i, j \in E} \text{Loss}(e_{i, j}, \log(1 + \exp(-U_i^T U_j))) \quad (9-70)$$

If there is social relation between user i and user j , then $e_{i, j} = 1$; otherwise $e_{i, j} = 0$. The implicit vector similarity between users i and j is calculated using cosine similarity $U_i^T U_j$, and $\text{Loss}(x, y)$ reflects the variance of the discrete binomial variable x and continuous variable y , and a loss function of the corresponding form can be used, such as square loss. The above formula requires that the user implicit vector obtained by matrix decomposition be consistent as much as possible with the actual relations existing between users.

Example 9-19 Matrix decomposition model SocialMF. Reference [91] proposes the SocialMF model, and the regular term is defined as the calculation result of the following formula:

$$L(U) = \sum_i \left\| U_i - \sum_{j \in N_i} T_{i, j} U_j \right\|_F^2 \quad (9-71)$$

Where, if user u and user v have relations, then matrix element $T_{u, v} = 1$; otherwise $T_{u, v} = 0$. The formula above requires that the vector expression obtained by neighbor users be as close as possible to the vector expression of the user itself.

To make the matrix decomposition model significant in probability, references [91 and

108] presented the corresponding probability generation model to model the user-item score matrix.

Take reference [91] as an example, suppose the implicit vectors of both users and items comply with Gaussian distribution, and the implicit vector expression of each user is influenced by its neighbor, then the posterior probability of the characteristic vectors of users and items can be expressed as the product of the likelihood probability function of score matrix and the priori probability function of the characteristic vectors, as indicated by the formula below:

$$\begin{aligned}
 p(\mathbf{U}, \mathbf{V} | \mathbf{R}, \mathbf{T}, \sigma_R^2, \sigma_T^2, \sigma_U^2, \sigma_V^2) &\propto p(\mathbf{R} | \mathbf{U}, \mathbf{V}, \sigma_R^2) p(\mathbf{U} | \mathbf{T}, \sigma_U^2, \sigma_T^2) p(\mathbf{V} | \sigma_V^2) \\
 &= \prod_{u=1}^N \prod_{i=1}^M \left[N(R_{u,i} | g(\mathbf{U}_u^T \mathbf{V}_i), \sigma_R^2) \right]^{I_{u,i}^R} \times \prod_{u=1}^N N\left(\mathbf{U}_u \middle| \sum_{v \in N_u} \mathbf{T}_{u,v} \mathbf{U}_v, \sigma_T^2 I\right) \\
 &\quad \times \prod_{u=1}^N N(\mathbf{U}_u | 0, \sigma_U^2 I) \times \prod_{i=1}^M N(\mathbf{V}_i | 0, \sigma_V^2 I)
 \end{aligned} \quad (9-72)$$

$N(x|\mu, \sigma^2)$ represents the Gaussian distribution with expectation as μ and variance as σ . The generation process is expressed using a probabilistic graphical model, as shown in Figure 9-8.

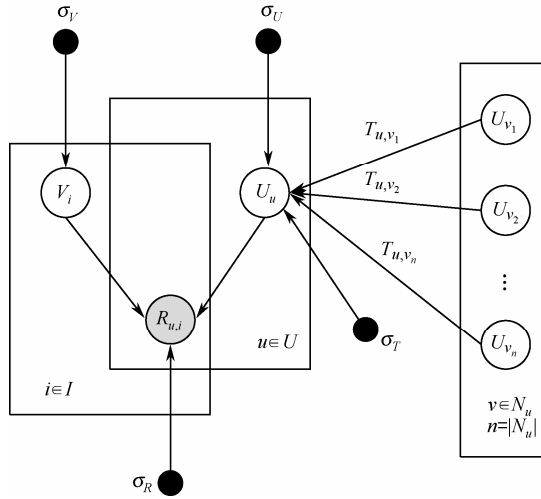


Figure 9-8 Graphical expression of Social MF model

In addition, some other research work also tried to perform social recommendation by establishing models. Reference [84] proposed a new matrix decomposition method, to solve the issue of social influence at the item level. Reference [109] has built a user interest spreading model based on friend relation, to

explain the user friend relations and interaction relations through random walking on the interest network of users, and thereby combining the issues of recommendation and linking prediction into one.

All the above work uses the known social network relations; however, under many circumstances, the relations between users cannot be directly observed. To solve this problem, many researchers have used the score similarity of users to simulate user relations. Reference [110] directly uses the user score similarity as degree of trust, and uses the trust transfer method to expand the closest neighbors.

Example 9-20 Metering of trust degree based on user score variance: references [111 and 112] proposed the metering method of trust degree based on user score variance, as shown by the following formula:

$$u(a, b) = \frac{1}{|R_a \cap R_b|} \sum_{i \in (R_a \cap R_b)} \left(\frac{|r_{a,i} - r_{b,i}|}{\max(r)} \right) \quad (9-73)$$

This method calculates the degree of trust between users by summing up the absolute values of errors of user a and user b on all score sets.

Reference [93] proposed the method to calculate according to score error proportions, which first defines the correctness evaluation method for scores between users, to classify the scores of users for the same item into a binary issue of being correct and incorrect by setting the threshold of errors. Reference [113] expanded this method, and proposed the non-binary judging method, to adjust the effect of score variance on user relations by introducing error panelizing factor. Reference [114] proposed the model of merging trust degree and similarity, which has the advantage to make use of trust transfer and most neighboring method.

In addition, references [89 and 115] proposed to improve the recommendation effect by using the user similarity and item similarity as the implicit social relations, use the Pearson's correlation coefficient to calculate user and item similarity, and add the similarity into the matrix decomposition model as a regularized parameter.

Reference [116] put forth the MCCRF model based on the condition random field, to establish links for user's scores; the conditional probability between different scores is estimated by using the user score similarity and item similarity, and the parameters of the model are trained by using the MCMC method.

9.5 Summary

This chapter presents the research development of information retrieval in social networks based on the three typical applications of search, classification and recommendation. In terms of search, we mainly summarize the current work from three aspects: query representation, document representation and similarity calculation. In terms of classification, we mainly summarize the progress of short text classification from the two dimensions of features and algorithms. In terms of recommendation, we mainly present how to add social information into traditional recommendation from the memory based and model based perspectives, to increase the effect of collaborative recommendation.

As shown from this chapter, information retrieval in social networks mainly involves short text representation and calculation. As the texts are short, it is an extensive practice to expand texts when they are expressed. The fact that social networks are rich in information makes it is an interesting topic to explore the ways of making use of such information, especially social networking information to improve the effect of text representation. In terms of calculation models, usually a model that can merge multiple features is used to introduce the features of microblogs.

We believe that the future development trend of information retrieval can be summarized as follows:

(1) Integrating more social network specific information: for example, information such as age, gender and geographic location of users in social networks has been proved useful in auto-completion query tasks^[117]. Such information can be completely used to retrieve social network information, and thereby realize personalized precise retrieval. Another example is that the dialogue context, social relations and even communities in the social network content can be very helpful for retrieval.

(2) More precise representations and computing models: as stated in this chapter, representation and calculation of short texts is the core content of information retrieval in the context of social network texts; therefore, in-depth study of such a core content is undoubtedly an important direction in the research of social network text retrieval.

(3) In-depth understanding of short text: researches on short text matching under the semantic space have been carried out recently^[118], therefore, how to mine the profound meanings of short texts will be an direction worth our attention.

(4) Novelty and diversity: in addition to correlation, novelty and diversity should also

be taken into account in social network information retrieval, in order to find diversified new information.

References

- [1] Christopher Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval[M]. Cambridge University Press. 2008.
- [2] Charu Aggarwal, Social Network Data Analytics[M]. Springer, March, 2011.
- [3] HP Luhn. A statistical approach to mechanized encoding and searching of literary information[J]. IBM Journal, 1957: 309-317.
- [4] Gerard Salton, Andrew Wong, ChungShu Yang. A vector space model for automatic indexing[J]. Communications of ACM , 1975: 613-620.
- [5] Gerard Salton, Christopher Buckley. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.
- [6] Amit Singhal, Christopher Buckley, and Mandar Mitra. Pivoted document length normalization[C]. In Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 1996: 21-29.
- [7] M. E. Maron, J. L. Kuhns. On relevance, probabilistic indexing and information retrieval[J]. Journal of the ACM, 1960: 216-244.
- [8] Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu and Mike Gatford. Okapi at TREC-3[C]. In Proceedings of the Third Text REtrieval Conference (TREC). Gaithersburg, USA, November 1994.
- [9] James Callan, Bruce Croft, Stephen Harding. The INQUERY retrieval system[C]. In Proceedings of 3th international conference on Database and Expert Systems Applications, 1992: 78-83.
- [10] Fabio Crestani, Mounia Lalmas, Cornelis J. Van Rijsbergen, I Campbell. Is this document relevant?... probably: a survey of probabilistic models in information retrieval[J]. ACM Computing Surveys (CSUR), 1998, 30 (4): 528-552.
- [11] Christopher D. Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing[M]. MIT Press, 1999.
- [12] Huang Changning, What is the main Chinese language information processing technology [N]. China Computer, 2002, 24.
- [13] Jay M. Ponte, Bruce Croft. A language modeling approach to information retrieval[C]. In Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval, New York, USA. 1998.

- [14] Chengxiang Zhai and John Lafferty, Model-based feedback in the language modeling approach to information retrieval[C]. In Proceedings of the 10th ACM international conference on Information and Knowledge Management (CIKM), pages 403-410, 2001. Atlanta, Georgia, USA.
- [15] Adam Berger and John Lafferty. Information retrieval as statistical translation[C]. In Proceeding of the 22rd international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), 1999: 222-229.
- [16] J J Rocchio. Relevance Feedback in Information Retrieval[R]. Prentice-Hall, 1975: 313-323.
- [17] Victor Lavrenko and Bruce Croft, Relevance based language models[C]. In Proceedings of the 24th international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), pages 120-127, 2001. New Orleans, Louisiana, USA.
- [18] Wouter Weerkamp, KrisztianBalog, and Maarten de Rijke, A Generative Blog Post Retrieval Model that Uses Query Expansion based on External Collections[C].In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP), 2009.
- [19] Miles Efron, Hashtag Retrieval in a Microblogging Environment[C]. In Proceeding of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), 2010.
- [20] Cunhui Shi, Kejiang Ren, Hongfei Lin, Shaowu Zhang, DUTIR at TREC 2011 Microblog Track.
- [21] Jaime Teevan, Daniel Ramage, Meredith Ringel Morris. #TwitterSearch: A Comparison of Microblog Search and Web Search[C]. In Proceedings of Web Search and Data Mining (WSDM), 2011: 9-12.
- [22] Xiaoyan Li and W. Bruce Croft, Time-based language models[C]. In Proceedings of the 12th ACM international conference on Information and Knowledge Management (CIKM), 2003: 469-475.
- [23] Wei Bingjie, Wang Bin, Combing Cluster and Temporal Information for Microblog Search [J]. Journal of Chinese Information Processing, 2013.
- [24] Giuseppe Amodeo, Giambattista Amati, and Giorgio Gambosi, On relevance, time and query expansion[C]. In Proceedings of the 20th ACM international conference on Information and Knowledge Management (CIKM), pages 1973-1976, 2011, Glasgow, Scotland, UK.
- [25] Mostafa Keikha, Shima Gerani, and Fabio Crestani, Time-based relevance models[C]. In Proceedings of the 34th international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), 2011: 1087-1088.
- [26] Stewart Whiting, Iraklis Klampanos, and Joemon Jose, Temporal pseudo-relevance feedback in microblog retrieval[C]. In Proceedings of the 34th European conference on Advances in Information Retrieval (ECIR), 2012.
- [27] Miles Efron, Peter Organisciak, and Katrina Fenlon, Improving retrieval of short texts through

- document expansion[C]. In Proceedings of the 35th international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), 2012: 911-920.
- [28] Yubin Kim, Reyyan Yeniterzi and Jamie Callan, Overcoming Vocabulary Limitations in Twitter Microblogs[C]. In Proceedings of the Twenty-First Text REtrieval Conference (TREC), 2012.
- [29] Li Rui, Wang Bin. Microblog Retrieval via Author Based Microblog Expansion [J]. Journal of Chinese Information Processing, 2014.
- [30] Alexander Kotov, Eugene Agichtein: The importance of being socially-savvy: quantifying the influence of social networks on microblog retrieval[C]. In Proceedings of the 22nd ACM international conference on Information and Knowledge Management (CIKM), 2013: 1905-1908.
- [31] Kamran Massoudi, Manos Tsagkias, Maarten Rijke, and Wouter Weerkamp. Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts[C]. in Advances in Information Retrieval, P. Clough, et al., Editors, 2011: 362-367.
- [32] Paul Ferguson, Neil O'Hare, James Lanagan, Owen Phelan, and Kevin McCarthy. An Investigation of Term Weighting Approaches for Microblog Retrieval[C]. in Advances in Information Retrieval, R. Baeza-Yates, et al., Editors, 2012: 552-555.
- [33] Xiaowen Li, Peiquan Jin, Xujian Zhao, Hong Chen, and Lihua Yue, NTLM: A Time-Enhanced Language Model Based Ranking Approach for Web Search Web Information Systems Engineering[C]. WISE 2010 Workshops, D. Chiu, et al., Editors, 2011: 156-170.
- [34] Nattiya Kanhabua and Kjetil Nørvåg. Using Temporal Language Models for Document Dating Machine Learning and Knowledge Discovery in Databases[C]. W. Buntine, et al., Editors, 2009: 738-741.
- [35] Miles Efron and Gene Golovchinsky, Estimation methods for ranking recent information[C]. In Proceedings of the 34th international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), 2011: 495-504.
- [36] BP Sharifi. Automatic Microblog Classification and Summarization[D]. University of Colorado, 2010.
- [37] Carter Simon, Tsagkias Manos and Weerkamp Wouter (2011) Twitter Hashtags: Joint Translation and Clustering[C]. In Proceedings of the ACM WebSci'11, 2011:14-17.
- [38] Beaux Sharifi, Hutton M. A, Kalita J.K. Experiments in Microblog Summarization[C]. In Proceedings of IEEE Second International conference on Social Computing (SocialCom), 2010.
- [39] Gustavo Laboreiro, Luís Sarmiento, Jorge Teixeira and Eugénio Oliveira. Tokenizing micro-blogging messages using a text classification approach[C]. In Proceedings of the fourth workshop on Analytics for noisy unstructured text data, 2010.
- [40] Daniel Ramage, Susan Dumais, Daniel Liebling. Characterizing Microblogs with Topic Models[C].

In Proceedings of ICWSM, 2010.

- [41] Einat Amitay, David Carmel, Michael Herscovici, Ronny Lempel and Aya Soffer. Trend detection through temporal link analysis[J]. Journal of the American Society for Information Science and Technology (JASIST), 2004, 55(14): 1270-1281.
- [42] Philip Yu, Xin Li and Bing Liu. On the temporal dimension of search[C]. In Proceedings of the 13th international World Wide Web conference on Alternate track papers posters, 2004: 448-449.
- [43] Klaus Berberich, Michalis Vazirgiannis, and Gerhard Weikum, Time-Aware Authority Ranking[J]. Internet Mathematics, 2005, 2(3): 301-332.
- [44] Jing Wan and Sixue Bai. An improvement of PageRank algorithm based on the time-activity-curve[C]. In Proceedings of IEEE International conference on Granular Computing(GRC), 2009.
- [45] Rinkesh Nagmoti, Ankur Teredesai, and Martine De Cock. Ranking Approaches for Microblog Search[C]. In Proceedings of 2010 IEEE/WIC/ACM International conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010.
- [46] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets[C]. In the 23th International Conference on Computational Linguistics (COLING), 2010: 295-303.
- [47] Xin Zhang Ben He, Tiejian Luo, and Baobin Li. Query-biased learning to rank for real-time Twitter search[C]. In Proceedings of the 21st ACM international conference on Information and Knowledge Management (CIKM), 2012:1915-1919.
- [48] Xinruo Sun, Haofen Wang, Yong Yu. Towards effective short text deep classification[C]. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 2011: 1143-1144.
- [49] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis[C]. In Proceedings of 22nd AAAI Conference on Artificial Intelligence (AAAI), 2007: 6-12.
- [50] Xiang Wang, Ruhua Chen, Yan Jia, Bin Zhou, Short Text Classification using Wikipedia Concept based Document Representation[C]. In Proceedings of International conference on Information Technology and Applications (ITA), 2013.
- [51] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]. In Proceedings of the 17th International World Wide Web Conference (WWW), 2008: 91-100.
- [52] Mengen Chen, Xiaoming Jin, and Dou Shen. Short text classification improved by learning multi-granularity topics[C]. In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI) , 2011: 1776-1781.

- [53] Sheila Kinsella, Alexandre Passant, and John G. Breslin. Topic classification in social media using metadata from hyperlinked objects[C]. In Proceedings of the 33th European conference on Advances in Information Retrieval (ECIR), 2011: 201-206.
- [54] Evgeniy Gabrilovich, Shaul Markovitch. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge[C]. In Proceedings of the 21st National Conference on Artificial Intelligence (NCAI), 2006: 1301-1306.
- [55] Duan Yajuan, Wei Furu, Zhou Ming. Graph-based collective classification for tweets[C]. .Proceedings of the 21st ACM international conference on Information and Knowledge Management (CIKM), 2012: 2323-2326.
- [56] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in Twitter to improve information filtering[C]. In Proceedings of SIGIR , 2010: 841-842.
- [57] Aixi Sun. Short Text Classification Using Very Few Words[C]. In Proceedings of SIGIR2012, 2012: 1145-1146.
- [58] Zitao Liu, Wenchao Yu, Wei Chen, Shuran Wang and Fengyi Wu. (2010). Short Text Feature Selection for Micro-Blog Mining[C]. In Proceedings of the International conference on Computational Intelligence and Software Engineering, 2010: 1-4.
- [59] Danesh Irani, Steve Webb. Study of Trend-Stuffing on Twitter through Text Classification[C]. In Proceedings of CEAS 2010, July 13-14: 114-123.
- [60] Quan Yuan, Gao Cong, and Nadia Magnenat Thalmann. Enhancing naive bayes with various smoothing methods for short text classification[C]. In Proceedings of the 21st International World Wide Web Conference (WWW), 2012: 645-646.
- [61] Mehran Sahami, Timothy D. Heilman, A web-based kernel function for measuring the similarity of short text snippets[C]. In Proceedings of the 15th International World Wide Web Conference (WWW), May 23-26, pages 377-386, 2006.
- [62] Sarah Zelikovitz, Haym Hirsh, Improving short-text classification using unlabeled background knowledge to assess document similarity[C]. In Proceedings of the 17th International conference on Machine Learning (ICML), 2000: 1191-1198.
- [63] Thiago Salles, Leonardo Rocha, Gisele Pappa, et al. Temporally-aware algorithms for document classification[C]. In Proceedings of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval(SIGIR), 2010: 307-314.
- [64] Kyosuke Nishida, Takahide Hoshide and Ko Fujimura. Improving Tweet Stream Classification by Detecting Changes in Word Probability[C]. In Proceedings of SIGIR2012, 2012: 971-980.
- [65] Choon Hui Teo and S. V. N. Vishwanathan. Fast and space efficient string kernels using suffix

- arrays[C]. In Proceedings of the 23rd International conference on Machine Learning (ICML), 2006: 929-936.
- [66] D. Okanohara and J. ichi Tsujii. Text categorization with all substring features[C]. In Proceedings of SDM, 2009: 838-846.
- [67] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu, Co-Clustering Based Classification for Out- of-Domain Documents[C]. In Proceedings 13th ACM SIGKDD international conference Knowledge Discovery and Data Mining (SIGKDD), Aug. 2007.
- [68] Rajat Raina, Andrew Y. Ng, and Daphne Koller, Constructing Informative Priors Using Transfer Learning[C]. In Proceedings of the 23rd International conference on Machine Learning (ICML), 2006: 713-720.
- [69] Pengcheng Wu and Thomas Dietterich. Improving SVM Accuracy by Training on Auxiliary Data Sources[C]. In Proceedings 21st International conference Machine Learning (ICML), July 2004.
- [70] Andrew Arnold, Ramesh Nallapati, and William W. Cohen, A Comparative Study of Methods for Transductive Transfer Learning[C]. In Proceedings of the Seventh IEEE International conference on Data Mining Workshops, 2007: 77-82.
- [71] Xiao Ling, Gui-Rong Xue, Wenyuan Dai, Yun Jiang, Qiang Yang, and Yong Yu, Can Chinese Web Pages be Classified with English Data Source[C]. In Proceedings of the 17th International conference on World Wide Web (WWW), 2008: 969-978.
- [72] Sinno Jialin Pan, Vincent Wenchen Zheng, Qiang Yang and Derek Hao Hu, Transfer Learning for WiFi-Based Indoor Localization[C]. in Workshop Transfer Learning for Complex Task of the 23rd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence, July 2008.
- [73] Dan Zhang, Yan Liu, Richard D. Lawrence, and Vijil Chenthamarakshan. Transfer Latent Semantic Learning?: Microblog Mining with Less Supervision[C]. In Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI), 2011: 561-566.
- [74] Guodong Long, Ling Chen, Xingquan Zhu, and Chengqi Zhang. TCSST: Transfer Classification of Short & Sparse Text Using External Data Categories and Subject Descriptors[C]. In Proceedings of the 21st ACM international conference on Information and Knowledge Management (CIKM), 2012: 764-772.
- [75] Paolo Avesani, Paolo Massa, and Roberto Tiella. A trust-enhanced recommender system application: Moleskiing[C]. In Proceedings of the 2005 ACM symposium on applied computing (SAC), 2005.
- [76] Paolo Massa, Paolo Avesani. Trust-aware recommender systems[C]. In Proceedings of the 2007 ACM conference on recommender systems(RecSys2007), 2007.
- [77] Matthew Richardson, Rakesh Agrawal, Pedro Domingos. Trust management for the semantic

- web[C]. The Semantic Web-ISWC 2003. Springer Berlin Heidelberg, 2003: 351-368.
- [78] R. Guha, Ravi Kumar, Prabhakar Raghavan and Andrew Tomkins. Propagation of trust and distrust[C]. In Proceedings of the 13th international conference on World Wide Web (WWW), 2004: 403-412.
- [79] Kailong Chen, Tianqi Chen, Guoqing Zheng, et al. Collaborative personalized tweet recommendation[C]. In Proceedings of the 35th international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), 2012: 661-670.
- [80] Jilin Chen, Rowan Nairn, Les Nelson, et al. Short and tweet: experiments on recommending content from information streams[C]. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2010: 1185-1194.
- [81] David Stern, Ralf Herbrich, Thore Graepel. Matchbox. Large Scale Online Bayesian Recommendations [C]. In Proceedings of the 18th International World Wide Web Conference (WWW), 2009.
- [82] <http://www.epinions.com/>.
- [83] Paolo Massa, Bobby Bhattacharjee. Using trust in recommender systems: an experimental analysis[J]. Trust Management. Springer Berlin Heidelberg, 2004: 221-235.
- [84] Peng Cui, Fei Wang, Shaowei Liu, et al. Who should share what?: item-level social influence prediction for users and posts ranking[C]. In Proceedings of the 34th international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), 2011: 185-194.
- [85] Ibrahim Uysal, Bruce Croft. User oriented tweet ranking: a filtering approach to microblogs[C]. In Proceedings of the 20th ACM international conference on Information and Knowledge Management (CIKM), 2011: 2261-2264.
- [86] Zi Yang, Jingyi Guo, Keke Cai et al. Understanding retweeting behaviors in social networks[C]. In Proceedings of the 19th ACM international conference on Information and Knowledge Management (CIKM), 2010: 1633-1636.
- [87] Maksims Volkovs, Richard Zemel. Collaborative Ranking With 17 Parameters[C]. In Proceedings of the 26th Neural Information Processing Systems (NIPS), 2012: 2303-2311.
- [88] Suhrid Balakrishnan, Sumit Chopra. Collaborative ranking[C]. In Proceedings of the fifth ACM international conference on Web Search and Data Mining(WSDM), 2012: 143-152.
- [89] Hao Ma. An experimental study on implicit social recommendation[C]. In Proceedings of the 36th international ACM SIGIR conference on Research and Development in Information Retrieval(SIGIR), 2013: 73-82.
- [90] Punam Bedi, Harmeet Kaur, Sudeep Marwaha. Trust Based Recommender System for Semantic Web[C]. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI),

2007: 2677-2682.

- [91] Mohsen Jamali, Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks[C]. In Proceedings of the fourth ACM conference on Recommender systems (RecSys), 2010: 135-142.
- [92] Sherrie Xiao, Izak Benbasat. The formation of trust and distrust in recommendation agents in repeated interactions: a process-tracing analysis[C]. In Proceedings of the 5th international conference on Electronic commerce, 2003: 287-293.
- [93] John O'Donovan, Barry Smyth. Trust in recommender systems[C]. In Proceedings of the 10th international conference on Intelligent user interfaces. ACM, 2005: 167-174.
- [94] Ioannis Konstas , Vassilios Stathopoulos , Joemon M Jose. On social networks and collaborative recommendation[C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), 2009: 195-202.
- [95] Paolo Massa, Paolo Avesani. Trust-aware collaborative filtering for recommender systems[C]. On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, 2004: 492-508.
- [96] Mohsen Jamali, Martin Ester. TrustWalker: a random walk model for combining trust-based and item-based recommendation[C]. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data mining (SIGKDD), 2009: 397-406.
- [97] Jennifer Ann Golbeck. Computing and Applying Trust in Web-based Social Networks[D]. University of Maryland College Park, 2005.
- [98] Cai-Nicolas Ziegler. Towards decentralized recommender systems[D]. University of Freiburg, 2005.
- [99] Hao Ma , Haixuan Yang , Michael R. Lyu , Irwin King. Sorec: social recommendation using probabilistic matrix factorization[C]. In Proceedings of the 17th ACM conference on Information and Knowledge Management (CIKM). ACM, 2008: 931-940.
- [100] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining (SIGKDD), 2008: 426-434.
- [101] Yehuda Koren. Factor in the neighbors: Scalable and accurate collaborative filtering[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2010, 4(1): 1.
- [102] Quan Yuan, Li Chen, Shiwang Zhao. Factorization vs. regularization: fusing heterogeneous social relationships in top-n recommendation[C]. In Proceedings of the fifth ACM conference on Recommender systems (RecSYS), 2011, 245-252.
- [103] Simon Funk. Netflix update: Try this at home[EB/OL]. <http://sifter.org/~simon/journal/20061211.html>, 2006.

- [104] Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering[C]. In Proceedings of KDD cup and workshop. 2007: 5-8.
- [105] Ruslan Salakhutdinov, Andriy Mnih. Probabilistic Matrix Factorization[C]. In Proceedings of the 21st Neural Information Processing Systems (NIPS), 2007, 1(1): 2.1.
- [106] Hao Ma, Irwin King, Michael R. Lyu. Learning to recommend with social trust ensemble[C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and Development in Information Retrieval(SIGIR), 2009: 203-210.
- [107] Hao Ma, Dengyong Zhou, Chao Liu, et al. Recommender systems with social regularization[C]. In Proceedings of the fourth ACM international conference on Web Search and Data Mining(WSDM). ACM, 2011: 287-296.
- [108] Yelong Shen, Ruoming Jin. Learning personal+ social latent factor model for social recommendation[C]. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1303-1311.
- [109] Shuang-Hong Yang, Bo Long, Alexander J. Smola, et al. Like like alike: joint friendship and interest propagation in social networks[C]. In Proceedings of the 20th international conference on World Wide Web (WWW), 2011: 537-546.
- [110] Manos Papagelis, Dimitris Plexousakis, Themistoklis Kutsuras. Alleviating the sparsity problem of collaborative filtering using trust inferences[J]. Trust Management. Springer Berlin Heidelberg, 2005: 224-239.
- [111] Georgios Pitsilis, Lindsay Marshall. A Model of Trust Derivation from Evidence for Use in Recommendation Systems[R]. In Technical Report Series, CS-TR-874. University of Newcastle Upon Tyne, 2004.
- [112] Georgios Pitsilis, Lindsay Marshall. Trust as a Key to Improving Recommendation Systems[J]. Trust Management, pages 210-223. Springer Berlin / Heidelberg, 2005.
- [113] Neal Lathia, Stephen Hailes, Licia Capra. Trust-Based Collaborative Filtering[C]. In Joint iTrust and PST Conferences on Privacy, Trust Management and Security (IFIPTM), Trondheim, Norway, 2008.
- [114] Chein-Shung Hwang, Yu-Pin Chen. Using trust in collaborative filtering recommendation[C]. New Trends in Applied Artificial Intelligence. Springer Berlin Heidelberg, 2007: 1052-1060.
- [115] Hao Ma, Irwin King, Michael R. Lyu. Learning to recommend with explicit and implicit social relations[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 29.
- [116] Xin Xin, Irwin King, Hongbo Deng, and Michael R. Lyu. A social recommendation framework based on multi-scale continuous conditional random fields[C]. In Proceedings of the 18th ACM conference on Information and Knowledge Management (CIKM), 2009: 1247-1256.

- [117] Milad Shokouhi, Learning to personalize query auto-completion[C]. In Proceedings of the 36th international ACM SIGIR conference on Research and Development in Information Retrieval(SIGIR): 103-112.
- [118] Zhengdong Lu, Hang Li: A Deep Architecture for Matching Short Texts[C]. In Proceedings of the 27th Neural Information Processing Systems (NIPS), 2013: 1367-1375.

The Rules of Information Diffusion in Social Networks

Thanks to their inherent liberality and openness, online social networks have gradually become an important information distributing center in contemporary society, and information diffusion in social networks has become unprecedentedly active. Studying the rules of information diffusion in social networks can help us to better understand the structure and attributes of online social networks as well as the rules associated with sudden events from the perspective of information diffusion. These research results have a wide range of applications in such fields as marketing, information pushing in shopping websites, public opinion monitoring and guidance, and so on. From the point of view of social benefits, social groups and government agencies can not only carry out information dissemination and improve the efficiency and transparency of management by drawing on the characteristics and laws of information diffusion, but also guide public opinions in a rational way by screening and filtering information based on such characteristics and laws. From the point of view of economic benefits, enterprises can optimize the allocation of resources according to the characteristics and laws of information diffusion in social networks, and thus facilitate the promotion and sales of products. Therefore, it is of great theoretical and practical significance to study the rules of information diffusion in social networks.

10.1 Introduction

Information Diffusion is a process in which people transmit, receive and feed back information by means of symbols and signals, for the purpose of achieving mutual understanding and influence by exchanging opinions, ideas and emotions. Information

diffusion in online social networks specifically refers to the diffusion of information realised via the medium of online social networks.

Information diffusion in online social networks has the following characteristics: first, the release and reception of information is extremely simple and fast; users are able to release and receive information through mobile phones and other mobile devices anytime, anywhere; secondly, information diffuses in the way of “nuclear fission”; as long as a message is released, it will be pushed by the system to all the followers, and once being forwarded, it will be spread out instantly among another group of followers. thirdly, everyone has the opportunity to become an opinion leader; each Internet user is allowed to play an important role in the occurrence, fermentation, dissemination and sensationalization of a sudden event; finally, it takes on the form of “We Media”; from super stars to grassroots, everyone can build their own “media”, where they can express themselves freely, create messages to the public, and communicate ideas, while receiving information from all directions on the platform. In general, social networks have sped up the process and expanded the breadth of information diffusion. As a result, people are able to access richer information through social networks.

Studying the law of information diffusion in social networks can help us to deepen our understanding of social networking systems and social phenomena, and thereby further understand the topologies, communication capabilities and dynamic behaviours of complex networks. In addition, the study of information diffusion is also conducive to researches in other aspects, such as, the discovery of models, identification of more influential nodes, and personalized recommendations.

The vigorous development of social networks provides a rich data base for researchers to carry out relevant researches, which allows them to study the mechanism of information diffusion and understand the law of information diffusion on the basis of massive real data, and have achieved phasic results. Information diffusion in social networks mainly involves the base network structure, the network users, the diffusing information, and other factors, and the relevant researches are also carried out around such factors. Research results based on the network structure mainly include the independent cascade model, the linear threshold model and their extended model. Research results based on the state of users mainly include the epidemic model and the influence diffusion model. Research results based on the characteristics of information mainly include the multi-source information diffusion model, and the competitive diffusion model. In view of the fact that explicit diffusion models cannot explain certain phenomena of information diffusion, some

researchers studied the method of information diffusion prediction based on certain given data, to predict the effect of information diffusion such as the popularity of hot topics in social networks. In view of the numerous and jumbled sources of information in social networks, some researchers studied the method of information source location, to search the original sources of information and track their diffusion paths based on the distribution status of the existing information, and thereby provide support for such applications as network security.

The sections of this chapter are organized as follows: Section 10.2 analyzes the main factors that affect the diffusion of information in social networks. Section 10.3, 10.4, and 10.5 describe the diffusion models and application examples based on network structure, group status and information characteristics, respectively. Section 10.6 introduces the method as well as the application examples of predicting the state of information diffusion based on certain given data. Section 10.7 describes the information source location method and analyses a number of cases. Finally, the challenges and prospects faced by the research work of information diffusion in social networks are proposed.

10.2 Influencing Factors Related to Information Diffusion in Social Networks

Based on mutual understanding, hobbies, personal worship or other factors, individuals in social networks connect with others to form a complex “structure of relationship”. Based on such a structure, the connected individuals gather together and form a “networking group” with common behavioral characteristics through mutual influence and interdependence. Based on the relational structure and social networking groups, all kinds of information is published and diffused. Therefore, the relational structure of online social networks provides a base platform for information diffusion, the social groups directly promote information diffusion, and the rich information serves as the necessary resources for information diffusion.

10.2.1 Structure of Social Networks

A social network is a social structure made up of a set of social actors and a set of ties

between these actors. From a macro perspective, different types of social networks can have different modes of information diffusion. For example, information in forums and blogs is mainly diffused in the “one-to-many” or “point-to-plane” mode. In these cases, after a disseminator releases a message, who will view it and how the viewers will react to it are unknown factors. There is not a definite connection between the transmitter and the receiver. In contrast, instant messaging services like QQ and MSN adopt the “point-to-point” communications methodology, where one can initiate a chat session with a particular individual whenever someone on your private list is online, and recipients are often defaulted to respond to this particular information; in the case of micro-blogging, it allows disseminators to disseminate information to their followers, and followers can choose to forward or comment on the information, or unfollow the object they used to follow and reject his information. As a result, one “node” can have connections with countless other nodes in the network, which ultimately brings forth a combination of various diffusion modes, such as, “one-to-multiple”, “one-to-one”, “multiple-to-multiple” and “multiple-to-one”.

In essence, it is the different strengths of the ties between nodes and the different densities of networking that gives rise to the diversified modes of information diffusion in online social networks. The strength of a tie is a combination of the connecting time, the emotional intensity, the intimacy, and the reciprocal services which characterize the tie ^[1]. It's usually defined as the relative overlap of the neighborhood of two nodes in the network. The greater the proportion of the two nodes' common neighborhood is, the greater the strength of the tie between the two nodes. Strong ties can lead to a closely tied and clustered community structure ^[2], with higher trust between nodes, which plays an important role in promoting individuals to reach a consensus^[3,4], while weak ties are usually formed between individuals who are not connected closely or frequently, and these individuals are usually dissimilar from one another; therefore, weak ties can provide new information, serving as sources of diverse information; thus, weak ties playing a more important role in wide-range information diffusion than strong ties ^[1,3,5,6]. The density of networking indicates the degree to which participants in the social network are interconnected. The closeness between individuals in a network reduces uncertainty and generates a sense of belonging, which may enhance trust between members and facilitate the diffusion of information ^[3,7].

10.2.2 Groups in Social Networks

From the perspective of network users, information diffusion is more or less

influenced by the different characteristics of users' behaviors. Mor Naaman et al. categorise network users into nine groups by analysing the contents of the status information from users on Twitter, of which the IS (information sharing), the OC (options/complaints), the RT (random thoughts) and the ME (about the user him/herself) categories account for the main part^[8]. Akshay Java et al. divide users into the following four categories according to users' intentions of using a social network: daily chatting, conversation, sharing information and news reporting. On this basis, according to the role of users in the diffusion of information, they are divided into the following three categories: information sources, information collection and friends^[9].

Normally, "celebrities" who have rich knowledge in certain areas or a wealth of personal experience may greatly enhance the trust of viewers in the information. Those people can always receive higher attention than ordinary people. Their messages are usually spread very fast and easily form an orientation of topics. These people are regarded as "information sources"; that is, the information released by them is of high credibility and value of dissemination. In the process of information diffusion, if the disseminator is an "information source", the information sent from him/her will be far more significantly diffused in terms of both strength and breadth than those of ordinary users.

Individuals in social networks grow into groups in the networks through aggregation and mutual influence. Compared with communication groups in reality, groups in online social networks are more interactive, open, cross-regional, and extensive. The tendencies of such groups are closely related to the diffusion of information. Mike Thelwall et al. evaluated the relationship between the development of popular topics and the emotional intensity of groups on Twitter by using the SetiStrength algorithm. The results show that popular events on Twitter are usually associated with an increase in the intensity of negative emotions, The intensity of positive emotions during the peak period is higher than that before the peak^[10].

10.2.3 Information

Different from social networks in the actual world which are formed on the basis of such factors as geographical location, common activities and kinship, users in online social networks communicate with each other and thus establish connections mainly through releasing, sharing, commenting and forwarding information. Therefore, information in online social networks carries all the records of users' online activities. The information

itself has such distinctive characteristics as timeliness, multi-source concurrency, subject diversity, etc., which plays an indispensable role in the analysis of information diffusion.

Multi-source of information means that users in a network acquire information not only through links in the online social network, but also through factors outside the network. For many traditional media or online media, online social networks can allow more users to access their high-quality information, which is an effective way for them to test new models of news report and explore new channels of communication; for online social networks, the participation of traditional media and online media brings over massive external information, which, coupled with the real social networks of their own, will safeguard the quality and scope of information diffusion.

Some messages, especially those in the same category, can have mutual influence when being diffused simultaneously in social networks, which distinguishes its rule of diffusion from those of independent information^[11]. In fact, the simultaneous diffusion of multiple inherently related messages is widely seen in social networks. For example, “Evergrande winning championship” and “Wang Feng’s confession”, “Red Cross” and “Guo Mei Mei”, “flu” and “Banlangen”, and so on. It is of great practical significance to study the multi-message diffusion mechanism in social networks.

10.3 Diffusion Model Based on Network Structure

This method is modelled mainly based on the structure of the network where information is diffused and the interaction between neighbor nodes. It is assumed that each node in the network has only two states: active or inactive. The active state indicates that the user receives a message; otherwise, it is inactive. Only the one-way change of state is considered; that is, from “inactive” to “active”. This book mainly introduces the linear threshold model, the independent cascade model and the extended model.

10.3.1 Linear Threshold Model

In 1978, Mark Granovetter conducted a research on the potential resistance of individuals whose participation in a collective activity was ascribed to the influence of their neighbours who had also participated in the same activity, and proposed a threshold model regarding collective behaviours^[12]. Drawing upon the ideas of such a threshold, researchers

have conducted extensive researches. Among them, the Linear Threshold Model is universally recognised.

In the Linear Threshold Model, each node v is assigned with a threshold $\theta(v) \in [0,1]$, representing its susceptibility to infections. Node w , which is adjacent to node v , influences node v with a nonnegative weight of $b_{v,w}$, and the sum of the $b_{v,w}$ values of all the w nodes neighboring node v is less than or equal to 1.

An inactive node v is activated only if the sum of the influence of its active neighbor nodes is greater than or equal to its threshold, as shown in Equation (10-1); i.e., the decision of an individual in the network is subject to the decisions of all its neighbor nodes, and the active neighbor nodes of node v can participate in the activation of v multiple times. Algorithm 10-1 shows the algorithm for implementing the linear threshold model.

$$\sum_{w: \text{active neighbor of } v} b_{v,w} \geq \theta_v \quad (10-1)$$

Algorithm 10-1 Linear Threshold Model Propagation Algorithm

- (1) Initial active node set A .
- (2) At time t , all the active neighbor nodes of node v attempt to activate v . If the sum of the influence of all active neighbor nodes exceeds the activation threshold of θ_v , node v is transitioned to the active state at time $t + 1$.
- (3) The above process is repeated until the sum of the influence of any active nodes already present in the network can not activate the neighbor node in the inactive state. Thus, the propagation process ends.

Example 10-1 Example of a Linear Threshold Model Application

In the network shown in Figure 10-1, we know that node a is the initial active node, and the thresholds of nodes b , c , d , e and f are 0.2, 0.4, 0.5, 0.6 and 0.2 respectively. The direction of the edge represents “being followed”; for example, $b \rightarrow c$ represents that b is followed by c ; that is, the message posted on the social network by b can be viewed by c , and b will have an effect on c . The weight of the edge indicates the size of the influence, where the influence of b on c is 0.3, and the range of influence is $[0,1]$.

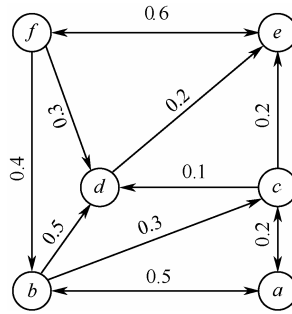


Figure 10-1 Example of a linear threshold model

The diffusion process based on linear threshold model is as follows in Figure 10-1:

Time Step 0: node a is activated.

Time Step 1: node a 's influence on node b is 0.5; node a 's influence on node c is 0.2. At this point, the influence on node b is 0.5, greater than its threshold of 0.2; thus node b is activated.

Time Step 2: node b 's influence on node c is 0.3; node b 's influence on node d is 0.5; node a 's influence on node c is 0.2. At this point, node c is subject to the influence of $0.3 + 0.2 = 0.5$, greater than its threshold of 0.4; thus, node c is activated.

Time Step 3: node c 's influence on node d is 0.1; node c 's influence on node e is 0.2; node b 's influence on node d is 0.5. At this point, node d is subject to the influence of $0.5 + 0.1 = 0.6$, greater than its threshold of 0.5; thus, node d is activated.

Time Step 4: node d 's influence on node e is 0.2, node c 's influence on node e is 0.2. At this point node e is subject to the influence of $0.2 + 0.2 = 0.4$, less than its threshold of 0.6. In this time step, no new node is activated; thus, the diffusion stops.

10.3.2 Independent Cascades Model

The Independent Cascades Model (IC) is a probabilistic model^[13,14] initially proposed by Jacob Goldenberg et al. in the research of a marketing model. The basic assumption of the model is that whether or not node u trying to activate its neighbor node v is successful is an event with a probability of $p_{u,v}$. And the probability that a node in an inactive state is activated by a neighbor node that has just entered an active state is independent of the activity of the neighbor who had previously tried to activate the node. In addition, the model also makes the assumption that any node u in the network has only one chance to attempt to activate its neighbor node v , whether or not it succeeds, and that even though

node u itself is still active at a later time, it does not have influence any more. Such a node is called non-influential active node. The implementation of the independent cascade model is described in Algorithm 10-2.

Algorithm 10-2 Independent Cascade Model Diffusion Algorithm

- (1) The initial active node set A .
- (2) At time t , when the newly activated node u attempts to influence its adjacent node v , the probability of success is $p_{u,v}$. If node v has multiple neighbor nodes that are newly activated, then these nodes will attempt to activate node v in any order.
- (3) If node v is activated successfully, then at time $t + 1$, node v becomes active and will have an effect on its adjacent non-active nodes; otherwise, node v has no change at time $t + 1$.
- (4) The process repeats until there is no influential active node exists in the network; thus, the diffusion process ends.

Example 10-2 Example of an independent cascade model application

The direction of the edge in Figure 10-2 represents “being followed”, and the weight of the node represents the probability of activation of the node. For example, the weight of $b \rightarrow c$ is 0.3, indicating that the probability of node b activating node c is 0.3.

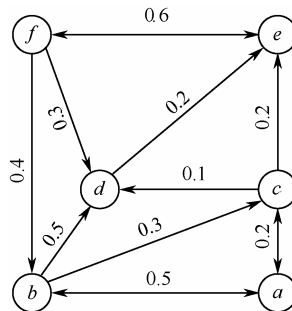


Figure 10-2 Example of an independent cascade model

The diffusion process of the independent cascade model based on Figure 10-2 is as follows:

Time Step 0: node a is activated.

Time Step 1: node a attempts to activate b with a probability of 0.5, and attempts to

activate c with a probability of 0.2, assuming that node b is successfully activated within this time step.

Time Step 2: node b attempts to activate c with a probability of 0.3 and attempts to activate d with a probability of 0.5, assuming that node c and node d are successfully activated within this time step.

Time Step 3: node c attempts to activate e with a probability of 0.2, node d attempts to activate e with a probability of 0.2, assuming that all the attempts within this time step have failed, and no new nodes are activated; thus, the diffusion ends.

10.3.3 Related Extended Models

No matter whether it is a linear threshold model or an independent cascade model, the diffusion process is simulated in a synchronous manner based on a discrete time axis, while in real social networks, the information diffuses along a continuous time axis and an asynchronous delay occurs in diffusion. Many researchers have improved the linear threshold and independent cascade models^[15~19]. Daniel Gruhl et al., by giving an algorithm for calculating the reading probabilities and replication probabilities between nodes, provided each edge with an occurrence probability of diffusion, allowing the independent cascade model to be applied to environments with potentially delayed diffusion^[15]. Kazumi Saito et al. used a continuous time and added a time delay parameter to each edge in the graph, extending the independent cascade model and the linear threshold model to the AsIC (Asynchronous Independent Cascades) and the AsLT (Asynchronous Linear Threshold) models^[18].

The above methods are focused on the reasoning of diffusion behaviors, without taking into account the influence of content on diffusion. Information diffusion is a complex social and psychological activity, and the content of diffusion inevitably affects the action between neighbor nodes. Wojciech Galuba et al. analysed the diffusion characteristics of URL in Twitter, starting from the attractiveness of URL, user's influence and the rate of diffusion, and constructed the linear threshold diffusion model of URL using these three parameters to predict which users would mention which URLs. The model can be used for personalized URL referrals, spam detection, etc.^[20]. Adrien Guille et al. proposed the T-BaSIC model (Time-Based Asynchronous Independent Cascades) using the Bayesian Logistic regression method on the basis of the AsIC model, from three dimensions: semantic meaning of topics, network structure and time, to predict the

probability of diffusion over time between nodes on Twitter. The experimental results show that the model has a good effect in predicting the dynamics of diffusion^[21].

To analyse and model the action between neighbor nodes by extracting the network structure based on the diffusion model of network topology is characteristic of simple model and easy expansion, which has certain advantages in the case of large social networks or limited information. In addition, such methods cross a number of disciplines such as graph theory, probability statistics, sociology and physics, and have a solid theoretical basis. The model is suitable for the study of diffusion cascade behavior, predict the diffusion path, and provide personalized recommendations according to the degree of the user accepting the information.

However, the topology-based diffusion model also has some drawbacks: first, from the perspective of timeliness, the social network topology that researchers achieved is static, equivalent to a snapshot of the original network, on which all explicit social relations before the acquisition are recorded; that is to say, all the connections established ten years ago and those at one second ago are collected at the same time; the connection that received just one notice and the connection of intimate communications between two friends are treated equally in the calculation model; secondly, in such a network topology, the weights of the connections are generally equal or identically distributed, meaning that the connected users have the same influence on each other, or that the influence among the users in the social network satisfies the simple probability function; thirdly, the role of other factors outside the network is ignored; users in the social network are affected not only by neighbor nodes, but also by the traditional media and the like in the external world, and thus participate in the diffusion of information. The next step should be focused on improving the existing methods based on the topology characteristics of dynamic networks, user influence analysis, external factors under consideration, and the introduction of corresponding parameters.

10.4 Diffusion Model Based on the States of Groups

There are two types of diffusion models based on group states: one is based on groups. By drawing upon the idea of the epidemic model, it categories nodes in a social network into several states, and depicts the process of information diffusion through changes of the states. The other is based on individuals, where a diffusion model based on the influences of individuals is established by taking into consideration the different roles of different

individuals on information diffusion.

10.4.1 Classical Epidemic Models

The most well-known theoretical model in the field of information diffusion is the epidemic model developed based on the spread of infectious diseases in reality. The epidemic model has a long history. As early as in 1760, Daniel Bernoulli used mathematical methods to study the spread of smallpox. At the beginning of the twentieth century, some scholars began to study the deterministic epidemic model. W.H.Hamer and Ronald Ross et al. made great contributions to the establishment of a mathematical model of infectious diseases. In 1927, William O. Kermack and Anderson G. McKendrick proposed the SIR model^[22] in the study of the Black Death in London and the plague in Mumbai. Taking into consideration the situation of repeated infection, they established the SIS model^[23] in 1932. Based on the study of these models, the “threshold theory” was proposed to distinguish the epidemic nature of diseases, which laid the foundation for the development of infectious disease dynamics. Michelle Girvan et al. added the concepts of immunity and mutation to explain the evolution of diseases^[24].

In the epidemic model, the individuals in a system are divided into several types; each type of individuals is in the same state. The basic states includes the Susceptible state (S); namely, healthy but susceptible to infection; Infected state (I), meaning it is infectious; the Recovered state (R), indicating that the infected body was cured and acquired immunity or was dead after infection. Different diffusion models are named after the transitions between such states. For example, if an individual shifts from the susceptible state into the infected state, such a diffusion model is named SI model; if a susceptible individual is infected, and becomes susceptible state again, such a diffusion model is called SIS model; if a susceptible individual is infected, but recovered and get immunity afterwards, such a diffusion model is named SIR model. The above mentioned models are introduced as follows.

1. SI Model

The SI model is used to describe diseases that cannot be cured after the infection, or infectious diseases that cannot be effectively controlled due to the emergent outbreak; for example, Black Death, SARS, etc. We may also say that in the SI model, once an individual is infected, it will remain in the infected state permanently. $S(i)$ and $I(j)$ are used to express

the susceptible population and the infected population respectively. Assuming that the individual becomes infected at the mean probability β , the infection mechanism can be expressed by Equation (10-2):

$$S(i) + I(j) \xrightarrow{\beta} I(i) + I(j) \quad (10-2)$$

At time t , the proportion of S-state individuals in the system is $s(t)$, and that of I-state individuals is $i(t)$. Based on the assumption, each infected individual can infect $\beta s(t)$ susceptible individuals. As the infected individuals have a proportion of $i(t)$, a total of $\beta i(t)s(t)$ susceptible individuals are infected. The dynamical model of the SI model can be described by the differential equations as shown in Equation (10-3):

$$\begin{cases} \frac{ds(t)}{dt} = -\beta i(t)s(t) \\ \frac{di(t)}{dt} = \beta i(t)s(t) \end{cases} \quad (10-3)$$

2. SIS Model

The SIS model is suitable for describing diseases such as colds and ulcers, which cannot be effectively immunized after cure. In the SIS diffusion model, the infected individuals, as the source of infection, transmit the infectious disease to susceptible individuals by a certain probability β , while the infected individuals return to the susceptible state with a certain probability γ . On the other hand, the susceptible individuals, once infected, becomes a new source of infection. The infection mechanism can be described by Equation (10-4):

$$\begin{cases} S(i) + I(j) \xrightarrow{\beta} I(i) + I(j) \\ I(i) \xrightarrow{\gamma} S(i) \end{cases} \quad (10-4)$$

Assuming that the proportion of S-state individuals in the system at time t is $s(t)$ and that of I-state individuals is $i(t)$, and that the growth rate of the infected individuals is $\beta i(t)s(t) - \gamma i(t)$ when susceptible individuals are fully mixed with infected individuals, the dynamic behavior of the SIS model can be expressed by the differential equations shown in Equation (10-5):

$$\begin{cases} \frac{ds(t)}{dt} = -\beta i(t)s(t) + \gamma i(t) \\ \frac{di(t)}{dt} = \beta i(t)s(t) - \gamma i(t) \end{cases} \quad (10-5)$$

3. SIR Model

The SIR model is suitable for diseases that, once being caught, can give the patients lifelong immunity, such as, smallpox and measles. Assuming that in unit time the infected individuals are in contact with some randomly selected individuals in all states at the average probability β , and recover and obtain the immunity at the average probability γ , the mechanism of the infection is as described in Equation (10-6):

$$\begin{cases} S(i) + I(j) \xrightarrow{\beta} I(i) + I(j) \\ I(i) \xrightarrow{\gamma} R(i) \end{cases} \quad (10-6)$$

Assuming that the proportions of individuals in the susceptible, the infected and the recovered states at time t in the system are $s(t)$, $i(t)$ and $r(t)$ respectively, and in the condition that the susceptible individuals are well mixed with the infected individuals, the growth rate of the infected individuals is $\beta i(t)s(t) - \gamma i(t)$, the decline rate of the susceptible individuals is $\beta i(t)s(t)$, and the growth rate of the recovered individuals is $\gamma i(t)$, then the dynamic behavior of the SIR model can be described as in Equation (10-7).

$$\begin{cases} \frac{ds(t)}{dt} = -\beta i(t)s(t) \\ \frac{di(t)}{dt} = \beta i(t)s(t) - \gamma i(t) \\ \frac{dr(t)}{dt} = \gamma i(t) \end{cases} \quad (10-7)$$

10.4.2 Infected Diffusion Models in Social Networks

The ideas of epidemic disease models are borrowed to divide nodes in a social network into the “susceptible” (S) group, to whom the information is still unknown, the “infected” (I) group, who have already received and keep transmitting the information, and the “recovered” (R) group, who have received the information but lost interest in transmitting it. The information diffusion is analysed based on the change of such different states^[25~27].

Example 10-3 Example of epidemic information diffusion model in social networks

Saeed Abdullah and Xindong Wu studied the diffusion of information on Twitter by using the SIR model^[25]. They believed that, similar to traditional epidemiology which takes

into account the birth rate, when the nodes in the infected state (Class I) in a social network tweet about something, the fans will become a new susceptible population, and the total number of them keeps growing (see Table 10-1). Assuming that the new susceptible population is introduced by the infected individual, they will establish a dynamic equation shown in Equation (10-8).

$$\begin{cases} \frac{dS}{dt} = -\beta \cdot S(t) \cdot I(t) + I(t) \cdot \mu \\ \frac{dI}{dt} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t) \\ \frac{dR}{dt} = \gamma \cdot I(t) \end{cases} \quad (10-8)$$

Table 10-1 Comparison of parameters in the epidemiology and Twitter dissemination model

	Epidemiology	Information dissemination on Twitter
$S(t)$	Susceptible individual set at time t	Set of users who can receive tweets from infected individuals at time t
$I(t)$	Infected individual set at time t	Set of individuals who tweet about certain topics at time t
$R(t)$	Recovered individual set at time t	Set of infected individuals who stopped tweeting about certain topics in a specific period of time
β	Infection rate	Diffusion rate
μ	Birth rate	The number of new fans that each infected individual gained in unit time
γ	Recovery rate	1/average infection time

The article focuses on three types of events:

(1) Twitter's internal events; i.e., the events appear and disappear on Twitter, whose external impact is limited. The "Follow Friday" event was selected.

(2) Real-time news; the "World Cup soccer match between the United States and Ghana on June 26, 2010" was selected.

(3) Social events; the "Memorial Day of the United States" was selected.

The experimental process of Saeed Abdullah et al.^[25] is as follows:

(1) Prepare the data set. For each event, a set of infected state $I(t)$, susceptible state $S(t)$ and recovered state $R(t)$ are maintained. Specific keywords are searched by using Twitter's API at an interval of Δt . Users who tweeted about a certain topic within $[t-\Delta t, t]$ are retrieved, to update the infected state set $I(t)$. The fans of each infected individual are retrieved; users already included in $I(t)$ are filtered out, and the remaining are added to the

susceptible state set $S(t)$. Users did not tweet about the topic in $[t-2\Delta t, t]$ are removed from $I(t)$ and added to the recovered state set $R(t)$.

(2) Computer simulation; three kinds of events are simulated by using the proposed model. The experimental results show that the model can effectively simulate the diffusion trend of events on Twitter.

10.4.3 Diffusion Models Based on Influence

Different individuals have different effects on information diffusion. For instance, authoritative users or users at a central position will have greater influence power to promote the diffusion of information.

Jaewon Yang et al. proposed a Linear Influence Model^[28] based on a large number of empirical studies on user behaviors on Twitter. The model assumes that the process of information diffusion is subject to the influence of certain individual nodes, and the trend of information diffusion is predictable by evaluating the influence of these nodes.

The influence function $I_u(l)$ of node u represents the number of fans referring to its message in a period of l after it was influenced. The function $v(t)$ represents the number of nodes in the system that refer to a message at time t . The LIM model assumes that $v(t)$ is the sum of the influence functions of all influenced nodes, as shown in Equation (10-9):

$$v(t+1) = \sum_{u \in A(t)} I_u(t - t_u) \quad (10-9)$$

Wherein $A(t)$ represents the set of influenced nodes, and node u is influenced at time t_u ($t_u \leq t$).

This model can be described as follows: nodes u , v , and w are influenced at time t_u , t_v and t_w respectively, after which each generates an influence function $I_u(t - t_u)$, $I_v(t - t_v)$ and $I_w(t - t_w)$. The quantity of a message being referred to in the system at time t , represented by $v(t)$, is the sum of these three influence functions.

Jaewon Yang et al. present the influence function of a node in a nonparametric way and estimate it by using the nonnegative least squares problem of the Mapping Newton Method^[29]. This model can effectively evaluate the influence of nodes and can be used to predict the dynamic changes of information diffusion over time.

The population-based diffusion model describes the dynamic changes in information diffusion by describing the state of acceptance of information by users in the network and the redistribution of individuals between these states. Such models are widely used in viral

marketing, rumor spread analysis, information source location and so on.

However, there are still some problems with the population-based diffusion model. In the epidemic diffusion model, the individuals are only classified in three states: infected, susceptible and immune, and each state will continue for some time, until the infection of the virus. However, in social networks, an individual's status after receiving a message is highly susceptible to the influence of the surrounding environment or other information, and such status also changes very fast. Based on the model of individual influence, because of the enormous scale and the huge number of nodes in a social network, and the fact that different opinion leaders can appear in different scenarios, it remains a major challenge to establish an influence-based diffusion model by identifying the key nodes, and estimating the influences of these nodes.

10.5 Diffusion Model Based on Information Characteristics

Information in online social networks carries all the records of users' online activities. The information itself has such distinctive characteristics as timeliness, multi-source concurrency, subject diversity, etc., which plays an indispensable role in the analysis of information diffusion. Different models can be created based on such characteristics.

10.5.1 Diffusion Analysis for Multiple Source Information

Most of the existing social network information diffusion models assume that information is not affected by factors outside the networks, and that it is only diffused among nodes along the edges of social networks. However, in the real world, users in social networks can access information through multiple channels.

When a node u in a social network releases information k and if none of its neighbor nodes have released information related to k , it indicates that u is influenced by some unobservable external factors, causing the unprecedented emergence of information k ; however, if its neighbor node has released the related information, then the fact that u releases information k may be the result of being influenced by its neighbors or some external factors.

Based on the above idea, Seth Myers et al. believe that in addition to acquiring information through network links, nodes in social networks also acquire information through external influences. Thereby, they created model^[30] as shown in Figure 10-3, in which function $\lambda_{\text{ext}}(t)$ is used to describe the amount of information that a user receives through external influences. If its neighboring nodes have posted relevant information, the user will have a link-based internal influence $\lambda_{\text{int}}(t)$ from them. Function $\eta(x)$ describes the probability that the user will post microblog after being exposed to the information. Eventually, the user will either post a relevant microblog under the influence, or cease reacting to the information.

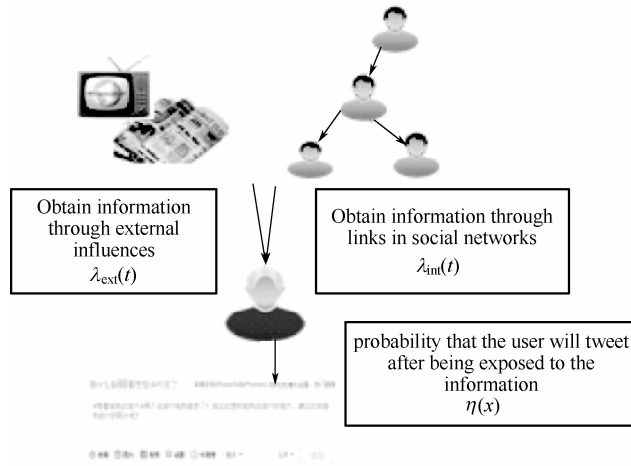


Figure 10-3 Schematic diagram of multi-source information influence

The total influence on node i is as shown in Equation (10-10):

$$P_{\text{exp}}^{(i)}(n; t) \approx \binom{t}{n} \left(\frac{A_{\text{int}}^{(i)}(t) A_{\text{ext}}(t)}{t} \cdot dt \right)^n \times \left(1 - \frac{A_{\text{int}}^{(i)}(t) + A_{\text{ext}}(t)}{t} \cdot dt \right)^{t/dt-n} \quad (10-10)$$

Wherein, $A_{\text{int}}^{(i)}(t)$ is the expected value of the node obtaining information through internal influence, and $A_{\text{ext}}(t)$ is the expected value of the node obtaining information through external influence.

Finally, the probability that user i will post microblogs after being exposed to the

information is as shown in Equation (10-11):

$$\begin{aligned}
 F^{(i)}(t) &= \sum_{n=1}^{\infty} P[i \text{ has } n \text{ exp.}] \times P[i \text{ inf.} | i \text{ has } n \text{ exp.}] \\
 &= \sum_{n=1}^{\infty} P_{\text{exp}}^{(i)}(n; t) \times \left[1 - \prod_{k=1}^n [1 - \eta(k)] \right]
 \end{aligned} \tag{10-11}$$

Wherein, function $\eta(x)$ describes the possibility that the user will post a microblog after having read a message.

Seth Myers et al. estimated the parameters of the model by using artificial networks and the infection time of some nodes. After applying this model to Twitter, they found that 71% of the information on Twitter was diffused based on the internal influence of the network, and the remaining 29% was triggered by factors outside the network.

10.5.2 Competitive Diffusion of Information

Some information, especially those in the same domain, can have mutual influence when spreading in a social network simultaneously; therefore, the diffusion rule of such information is different from that of independent information. To address the case of multi-source information diffusion, Alex Beutel et al. introduced the interaction factor to describe the intensity of interaction between two messages. The SI1|2S model^[31] featuring the mutual influence of information was proposed based on the expansion of the SIS epidemic model.

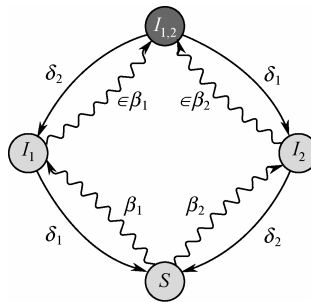


Figure 10-4 SI1|2S model's four states transition diagram^[31]

The model assumes that the nodes have four states: I_{12} indicates that the node is infected with both virus 1 and virus 2; I_1 indicates that the node is only infected with virus

1; I_2 indicates that the node is only infected with virus 2; S indicates that the node is not infected. The cure rate is δ , representing the probability of recovery from the infected state. The cure rates of virus 1 and virus 2 are δ_1 and δ_2 respectively. The infection rate is β ; the node in the S state is infected with virus 1 at the probability of β_1 , and infected with virus 2 at the probability of β_2 . The node infected with virus 1 has a probability of $\epsilon \beta_2$ to be infected with virus 2, and the node infected with virus 2 has a probability of $\epsilon \beta_1$ to be infected with virus 1, as shown in Figure 10-4.

Assuming that I_1 , I_2 and I_{12} indicate the number of nodes in each of the three states, the number of nodes in each state changes with time as follows:

$$\frac{dI_1}{dt} = \beta_1 S(I_1 + I_{12}) + \delta_2 I_{12} - \delta_1 I_1 - \epsilon \beta_2 I_1(I_2 + I_{12}) \quad (10-12)$$

$$\frac{dI_2}{dt} = \beta_2 S(I_2 + I_{12}) + \delta_1 I_{12} - \delta_2 I_2 - \epsilon \beta_1 I_2(I_1 + I_{12}) \quad (10-13)$$

$$\frac{dI_{12}}{dt} = \epsilon \beta_1 S_2(I_1 + I_{12}) + \epsilon \beta_2 S_1(I_2 + I_{12}) - (\delta_1 + \delta_2)I_{12} \quad (10-14)$$

If N is the total number of nodes, then the number of nodes in the S state is:

$$S = N - I_1 - I_2 - I_{12} \quad (10-15)$$

Threshold $\epsilon_{\text{critical}}$ satisfies Equation (10-16):

$$\epsilon_{\text{critical}} = \begin{cases} \frac{\sigma_1 - \sigma_2}{\sigma_2(\sigma_1 - 1)}, & \sigma_1 + \sigma_2 \geq 2 \\ \frac{2(1 + \sqrt{1 - \sigma_1\sigma_2})}{\sigma_1\sigma_2}, & \sigma_1 + \sigma_2 < 2 \end{cases} \quad (10-16)$$

Wherein, ϵ reflects the interaction between the two viruses: when $\epsilon > \epsilon_{\text{critical}}$, both viruses can coexist; when $\epsilon = 0$, both viruses are immune to each other; when $0 < \epsilon < 1$, both viruses compete with each other; when $\epsilon = 1$, both viruses do not affect each other; when $\epsilon > 1$, both viruses promote each other's diffusion.

Alex Beutel et al. selected two video service websites, Hulu and Blockbuster, and two browsers, Firefox and Google Chrome, as the cases for study. The search volume of the relevant information is obtained from Google Insights, and the data is fitted with SI1|2S model. This model can fit the data well, which shows the applicability of the model.

Multi-source information diffusion analysis, through the modeling of information

sources, helps us understand the mechanism of action between the real world and online social networks. In real scenarios, different information spreads through social networks at the same time; the research on the competitive diffusion of information helps to establish a model that can better reflect information diffusion in the real world, and thus improve our understanding of the law of diffusion.

The existing researches mainly start from the scope, time and other factors of information diffusion, and a number of results have been achieved. In addition to these factors, other properties of information such as time, content, and source constitute the inherent attribute of its diffusion. In what manner is the role of information combined with the role of users, when it is diffused in a social network? To investigate into this problem will help people to better understand the mechanism of diffusion.

10.6 Popularity Prediction Method

An information diffusion model describes the law of information diffusion, but how to express and measure the overall effect of information diffusion in social networks is also a question worthy of study. Some tweets posted by celebrities will be quickly forwarded and commented by their fans, and lead to hot discussions in society. For example, the message about the “divorce of Faye Wong and Li Yapeng” on Weibo was forwarded for more than a hundred thousand times in just half an hour, while messages posted by some users are rarely viewed. The number and frequency of certain behaviors of users in social networks are associated with the significant difference between the spreading of online contents in terms of pace and scope. In this book, popularity is used to measure the overall degree and effect of information diffusion.

Popularity in different forms of media is reflected in different ways. For example, the popularity of a BBS post can be the number of replies to the posts. The popularity curve of the post about the “missing of Malaysia Airlines flight” on Tianya Forum is shown in Figure 10-5; for a video clip, its popularity can be the number of views; the popularity of a message on Weibo can be the sum of forwards and responses. In general, the greater the popularity value of an online content is, the more “popular” the online content is; that is, the more widely or deeper it is spread. In this section, we introduce a couple of ways to predict the popularity of information diffused in social networks.

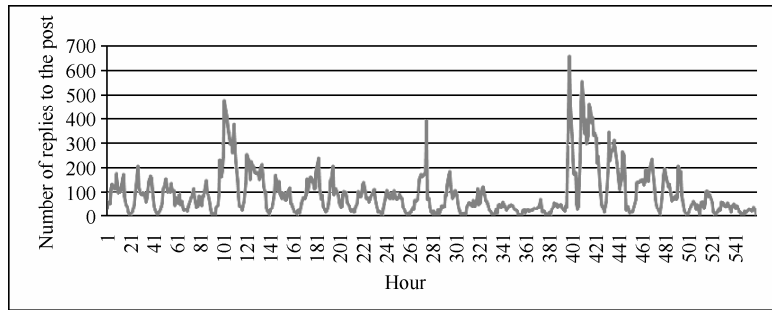


Figure 10-5 Popularity curve of the post about the “missing of Malaysia Airlines flight”

10.6.1 Prediction Models Based on Historical Popularity

Some researchers believe that there is a correlation between historical popularity and future popularity. A regression model that shows the correlation between historical and future popularities is established by considering the popularity at a particular point in the early stage or at a series of time points. The classical model is the SH model proposed by Gabor Szabo and Bernardo A. Huberman in 2008^[32]. The results were verified by posts on Digg and videos on YouTube.

In order to better observe the relationship between early popularity and late popularity, Gabor Szabo and Bernardo A. Huberman pretreated the data and found that if the popularity is logarithm transformed (\ln), the early popularity and late popularity will show a strong linear correlation, and the random fluctuations can be expressed in the form of additive noise. They also built a scatter plot to describe the early popularity and late popularity of the data set by means of mathematical statistics. As far as the Digg posts are concerned, each scattered point represents a post sample, and the abscissa value that each point corresponds to represents the number of “likes” for the post 1 hour after it was released, while the ordinate value represents the number of “likes” 30 days later. In the case of videos on Youtube, each scattered point represents a video sample, and the abscissa value that each point corresponds to represents the number of views 7 days after the video was posted, while the ordinate value represents the number of views 30 days later. According to the scatter plot, the specific relationship between the early popularity and late popularity of Digg posts and that of Youtube videos can be discovered by using the least squares method: $\ln y = \ln x + 5.92$, $\ln y = \ln x + 2.13$.

The SH model established according to the above process is as follows:

$$\ln N_s(t_2) = \ln r(t_1, t_2) + \ln N_s(t_1) + \varepsilon_s(t_1, t_2) \quad (10-17)$$

Wherein, $N_s(t_2)$ represents the popularity of an online content s at time t_2 , while $\ln N_s(t_2)$ is the dependent variable, indicating the late popularity. $\ln r(t_1, t_2)$ is the intercept. $\ln N_s(t_1)$ is an independent variable, indicating the early popularity. ε_s is the additive noise, i.e. the error term.

As far as linear regression fitting is concerned, if the error is subject to the normal distribution, then the fitting is correct. Therefore, the normality of the residual is detected by the quantile-quantile plot (QQ plot). It is found that the effect of the SH model for fitting the early popularity and the late popularity is acceptable. The QQ plot is used to visually verify whether a set of data is from a certain distribution, and is often used to check whether it is subject to the normal distribution.

10.6.2 Prediction Models Based on Network Structure

Peng Bao et al.^[33] improved the SH model, by taking into account the effect of the structural characteristics of a network on popularity. Using Sina Weibo as the object of study, they found that the structural characteristics of early microblog forwarders could reflect the final popularity of an online content.

The researchers first measured the correlation between the final popularity and the Link Density, and that between the final popularity and the Diffusion Depth of microblogs, and discovered a strong negative correlation between the final popularity and the link density, and a strong positive correlation between the final popularity and the diffusion depth. This shows that groups of low link density and high diffusion depth are more conducive to enhancing the popularity of microblogs. Based on the above findings, the researchers improved the SH model.

The improved model is shown in Equation (10-18):

$$\ln \hat{p}_k(t_r) = \alpha_1 \ln p_k(t_i) + \alpha_2 \ln \rho_k(t_i) + \alpha_3 \quad (10-18)$$

Wherein, $\hat{p}_k(t_r)$ is the popularity at time t_r , i.e., the late popularity; $p_k(t_i)$ is the popularity at time t_i , i.e., the early popularity; $\rho_k(t_i)$ is the link density at time t_i . $\alpha_1, \alpha_2, \alpha_3$ are the parameters trained from the data set.

$$\ln \hat{p}_k(t_r) = \beta_1 \ln p_k(t_i) + \beta_2 \ln d_k(t_i) + \beta_3 \quad (10-19)$$

Wherein, $\hat{p}_k(t_r)$ is the popularity at time t_r , i.e., the late popularity; $\ln p_k(t_i)$ is the

popularity at time t_i , i.e., the early popularity; $d_k(t_i)$ is the diffusion depth at time t_i . β_1 , β_2 , β_3 are the parameters trained from the data set (See Table 10-2).

Table 10-2 Comparison of network structure based model and SH model

Model	RMSE	MAE
SH model	0.77	0.57
Link density model	0.63	0.45
Diffusion depth model	0.61	0.43

The comparison of network structure based models and SH model is shown as Table 10-2. It can be seen from the table that the RMSE (Root Mean Squared Error) and the MAE (Mean Absolute Error) of the improved model are significantly reduced compared with the SH model.

10.6.3 Prediction Models Based on User Behaviors

Some researchers believe that the promotion of the popularity of online contents is closely related to the behaviors of social network users. Kristina Lerman and Tad Hogg argued that user's social behavior, such as, registering a website, reading a message, "liking" a message, becoming friends or fans of the message poster, etc. Such behaviours can be represented with the state transition in the Stochastic Process^[34]. At the same time, user behaviors can also determine the visibility of online contents. Take Digg for example. After having sufficient "likes", a post will be pushed to the home page, where its visibility is elevated. The higher the visibility of a post is, the more likely its popularity is enhanced, while the more hidden a post is, the more likely it stays invisible.

The model established by Kristina Lerman and Tad Hogg is as follows:

$$\frac{dN_{\text{vote}}(t)}{dt} = r(v_f(t) + v_u(t) + v_{\text{friends}}(t)) \quad (10-20)$$

Wherein, $N_{\text{vote}}(t)$ represents the number of "likes" a post receives at time t , i.e., the popularity. r represents the fun factor of the post, i.e., the probability of users liking it once after being viewed. v_f , v_u , v_{friends} respectively represent the rates of users seeing the post via the front page, the "upcoming" section, and the "friends" interface.

The researchers assume that all users will first browse the front page after visiting the Digg website, and then enter the "upcoming" section at a certain probability. The posts

posted by all users on Digg are grouped, with every 15 in one group. The latest 15 posts will be on the first page, the next 15 on the second page, and so on. Researchers use the function $f_{\text{page}}(p)$ to indicate the visibility of a post. If p has a value of 1.5, it means that the post is in the middle of the second page. $f_{\text{page}}(p)$ decreases as p increases, while p increases as time increases. Researchers use the Gaussian inverse function to represent the distribution of the number of pages viewed by users. Finally measure the v_{friends} , i.e., the rate of a post being viewed via the “friends” interface. Users can see the posts both submitted and liked by friends. Researchers use the function $s(t)$ to represent the number of friends who haven’t seen the post, out of the total number of the friends of the liker. Suppose a friend of the liker finally sees the post at the rate of w , then $v_{\text{friends}} = ws(t)$.

So here we have:

$$v_f = v f_{\text{page}}(p(t)) \Theta(N_{\text{vote}}(t) - h) \quad (10-21)$$

$$v_u = c v f_{\text{page}}(q(t)) \Theta(h - N_{\text{vote}}(t) \Theta(24hr - r)) \quad (10-22)$$

$$v_{\text{friends}} = ws(t) \quad (10-23)$$

Wherein, t is the length of time after the post was submitted, and v is the rate at which the user visits Digg.

Wherein, the rate of users visiting Digg, the probability of browsing the “upcoming” section, the rate of liker’s fans visiting Digg, the distribution of page views and some other parameters are the empiric values trained from the train set; the fun factors and the number of fans of the poster vary with different posts (see Table 10-3).

Table 10-3 Values of the parameters in the model

Parameter	Value
The frequency of users visiting Digg	$v = 10$ users / min
The probability of browsing the “upcoming” section	$C = 0.3$
The rate of liker’s fans visiting Digg	$\omega = 0.002$ / min
Distribution of page views	$\mu = 0.6, \lambda = 0.6$
The number of the liker’ fans	$a = 51, b = 0.62$
The number of likes required for pushing a content	$h = 40$
The update rate of posts in the “upcoming” section	$k_u = 0.06$ pages / min
The update rate of posts on the front page	$k_f = 0.003$ pages / min
Posts feature parameters	
Fan factor	r
Number of the poster’s fans	S

10.6.4 Prediction Models Based on Time Series

Many methods for predicting the popularity of the online contents are based on sample set; that is, to establish a mathematical model that describes the relationship between different factors and future popularity, and the model parameters are trained from the sample set. This kind of methods that are based on the sample set can accurately predict the long-term popularity of common online contents, but the accuracy of predicting the popularity of hot online contents is low. Because it is difficult to find a suitable sample set for hot online content, which will be verified at the end of this section. What makes a hotspot content hot is that it has some characteristics that distinguish it from the ordinary ones. It is not appropriate to collect ordinary contents into a sample set, nor select hotspot contents to create a sample set, because the popularity of each hotspot content has its own status. If the samples in a sample set do not share any common characters, such a sample set is a failure. Therefore, in this section we will describe a method for predicting the popularity of online contents using a time series based model^[35]. This method does not require a sample set, but simply the historical data of the online content. It allows analyzing the statistical laws and characteristics of the historical data.

1. Time Series Method

Time series is the series of the different values of an observed variable at different points in time arranged in the sequence of time. The basic assumption of the time series method is the notion of continuity in the development of things; that is, history provides a means to predict the future.

The time series consists of four constituent components: level, trend, seasonality, and noise, wherein “level” reflects the average level of the observed variables in this group; “trend” indicates the increases or decreases of the time series in a period of time; “seasonality” refers to the repeated fluctuations of the time series in a period of time, i.e., the aforementioned periodicity. The “level”, “trend” and “seasonality” are called the systematic part, while “noise” is the non-systematic part. Time series is aimed to make predications on the systematic part.

These different components together constitute the entire time series. The compound modes of different components are divided into addition and multiplication.

Addition mode: $Y_t = \text{level} + \text{trend} + \text{seasonality} + \text{noise}$

Multiplication mode: $Y_t = \text{level} \times \text{trend} \times \text{seasonality} \times \text{noise}$

The seasonality components of a time series are sub-divided into additive season and multiplicative season. The behavior of the additive season is that the seasonal fluctuation does not change with the overall trend and level of the time series. The behavior of the multiplicative season is that the seasonal fluctuation changes with the overall trend and level of the time series. The type of season selected for analysis determines whether to choose an additive model or a multiplication model. For time series of additive seasons, the additive model is usually selected for fitting, while for time series of multiplicative seasons, the multiplicative model is usually selected for fitting.

The basic steps of prediction using the time series method are as follows:

- (1) access to data;
- (2) visual analysis: analyse data characteristics, select the data granularity, and develop the time series plot;
- (3) data preprocessing: deal with missing values, extreme values, etc.;
- (4) data division: division of train set and validation set;
- (5) application of prediction methods: select the appropriate model;
- (6) evaluation of the prediction performance: use MAPE (Mean Absolute Percentage Error), RAE (Relative Absolute Error), MSE (Mean Square Error) and other methods for performance evaluation.

2. Time Series Models

There are some classical methods for prediction in time series, such as, regression and smoothing. This section chooses the typical additive MLR (Multiple Linear Regression Model) model in the regression method and the typical multiplication model HW (Holt-Winters) model in the smoothing method.

The time series in the MLR model is presented as follows:

$$Y_t = P_t + (a_1x_1 + a_2x_2 + \dots + a_{m-1}x_{m-1}) + E_t \quad (10-24)$$

Wherein, t represents time; Y represents the actual value of the time series; P_t is a polynomial, representing the trend and the level terms; m is the length of season, x_1, x_2, \dots, x_{m-1} are the dummy variables; in the case of m periods, there are $m-1$ dummy variables; the period without a corresponding dummy variable is the reference value. The dummy variable is either 0 or 1. If the time falls within a specific period, the dummy variable in this period is 1, and the others are 0. a_1, a_2, \dots, a_{m-1} are the coefficients corresponding to the dummy variables respectively. E_t represents noise.

The time series in the HW model is presented as follows:

$$Y_t = (L + tT)S_t + E_t \quad (10-25)$$

Wherein, t presents time, Y is the fitted value of the time series, L is the level component, T is the linear trend component, S is the seasonal factor, and E is the noise component

$$L_t = \alpha y_t / S_{t-m} + (1 - \alpha)(L_{t-1} + T_{t-1}) \quad (10-26)$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \quad (10-27)$$

$$S_t = \gamma(y_t / L_t) + (1 - \gamma)S_{t-m} \quad (10-28)$$

Wherein, y_t represents the observed value at time t , m is the length of the season, and α , β , and γ are called smoothing parameters, which can be calculated with the least MSE (Mean Squared Error) of the train set.

The multiplicative HW model performs exponential smoothing for level, trend, and seasonality components. The smoothing parameters determine the rate of the latest information. The more approximate to 1 the value is, the more recent new information is used.

Thus, the k -step ahead-of-time prediction can be achieved in Equation (10-29).

$$Y_{t+k} = (L_{t-1} + kT_{t-1})S_{t+k-m} \quad (10-29)$$

Example 10-4 Example of online content popularity prediction based on time series

The above sections depicted the basic idea of time series and two time series models. In the following part, we will describe the specific application of the time series approach in predicting the popularity of online contents - predicting the popularity of topics on Tianya Forum. Two types of hot topics were selected for analysis and prediction: one is the long-term, non-emergent type, such as, Christmas and the US general election; the other is the short-term, emergent type, such as, H7N9 and Beijing haze.

1) Data set collection

Source: Sina News Center (<http://news.sina.com.cn>) and Tianya Forum (<http://www.tianya.cn>). Sina News Center is one of the major channels of Sina, with 24-hour rolling coverage of domestic, international and social news, and more than 10,000 pieces of news issued daily. Tianya Forum is the largest forum website in China, founded in 1999, with registered users up to 85 million, and daily visits of about 30 million. The Forum consists of politics, entertainment, economy, emotion, fashion and other sections. As of June 2013, the posts in the entertainment section had exceeded 3.2 million, and the replies exceeded 170 million; the posts in the economy section has exceeded 1.4 million,

and the replies exceeded 17 million.

The data set was collected into two steps:

- (1) select hot topics from Sina News Center;
- (2) search on Tianya Forum for the hot topics screened out in the first step.

The popularity of hot topics is here defined as the quantity of all posts and replies related to the topic within a certain period of time.

Hot topics in Sina News Center starting from July 5, 2004 were collected. Topics on a specific date can be searched by date. 10,000 topics were collected from the international, social, entertainment, sports, science and technology, and other sections, which were then artificially divided into two types: emergent and non-emergent, with 6,000 emergent-type topics, and 4,000 non-emergent-type topics. After that, the topics collected in the first step were used as keywords, which were searched in Tianya Forum by using its internal searching section. There are multiple searching methods including searching by relevance, searching by posting time, searching from the full text, searching titles, and so on. We chose the relevance and the title searching methods. After further screening, we selected 7,000 hot topics, of which 3,000 are non-emergent and 4,000 are emergent topics. The non-emergent topics collected were data from January 2001 to December 2012, and the popularity of each topic is higher than 15,000; the emergent topics collected were data from January 2011 to December 2013, and the popularity of each topic in the first 15 days is higher than 18,000.

2) Analysis of characters of time series

Seasonality is also known as periodicity. Non-emergent topics tend to show seasonality when periodic events occur or are about to occur. Take the US general election for example, whenever the season of election arrives, the popularity of such a topic begins to rise, and the length of season stretches for 4 years, as shown in Figure 10-6. As for emergent topics, they are discussed by enormous people every day in a short period of time. In a day, the amount of discussions reaches a climax between 21:00 - 23:00, but drops to the lowest point between 2:00 - 4:00, which is in line with the habits of people. Therefore, the length of season is 24 hours. A topic goes through a process of emerging, climaxing, and declining. Hot topics show smaller fluctuations in the decline period than in the peak period. Therefore, the seasonality of the time series of the topic's popularity is multiplicative, which will be proved by the next experiment.

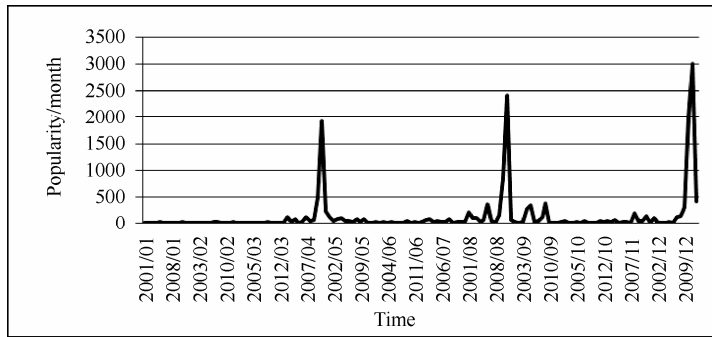


Figure 10-6 Popularity of the “US general election” topic from January 2001 to December 2012

If the data is too fine-grained, many values would result in 0. For example, non-emergent topics are not discussed much, and if the granularity is set at the level of “hour”, the popularity is calculated once an hour. As a result, many of the values might be 0, and it is not easy to identify the seasonal characteristics. Therefore, it’s better to select “month” as the granularity. Take the topic on “Christmas” for example, the value reaches the peak in December every year, while it is quite lower in the other months. This shows strong seasonality and the season length is 12 months; however, if the data granularity is too rough, some seasonal characteristics of a time series are likely to be overlooked. For example, an emergent topic becomes a hot topic among many people. If “day” is selected as the granularity, it is not easy to observe the period, but if “hour” is used as the granularity, the law of the hourly quantity of replies in every 24 hours is found to coincide with the habits of people, which is the characteristic of seasonality, with a season length of 24 hours. Figure 10-7 shows the frequency of replies on Tianya Forum in each time period of a day. The statistics is based on 35,619,985 replies to 10,000 posts on Tianya Forum.

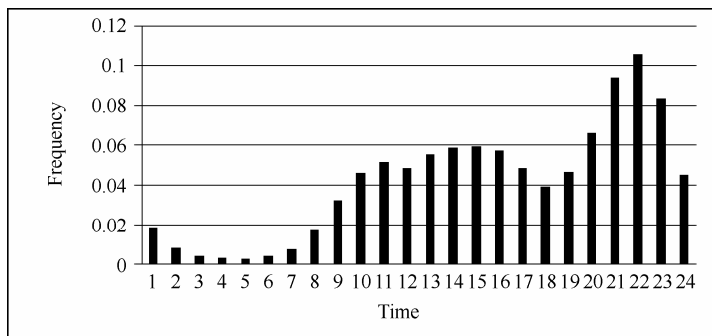


Figure 10-7 Frequency distribution of replies to hot posts on Tianya BBS in different periods of a day

3) Prediction experiment and the results

We chose “Beijing haze” as an example of emergent type topics, and “Christmas” as an example of non-emergent topics.

Figure 10-8 shows the time series of the popularity of the “Beijing haze” topic. The topic’s popularities respectively on the 5th, 6th and 7th day after the occurrence were chosen to form a train set, to predict the popularity on the 8th day, i.e., the peak day, which served as a validation set. In MLR model here, a quadratic polynomial with one variable is used as the trend; because the data in the train set is limited, a polynomial of higher order may lead to over-fitting. The multiplicative HW model minimizes the MSE of the train set. α, β, γ are determined, where $\alpha = 0.98, \beta = 0.54, \gamma = 0.01$. The results are shown in Figure 10-9. The time series predicted by the MLR model train set is negative in the location of lower trend, because the MLR is an additive model. The seasonal fluctuation is stable, perhaps because the selected trend terms are lower than the actual ones. The accuracy rate of the MLR model in predicting the popularity on the 8th day is 79.1%, and the accuracy rate of the HW model is 88.3%.

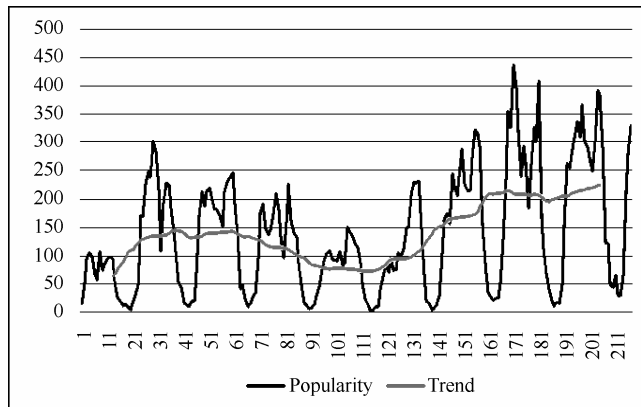


Figure 10-8 The blue line indicates the time series of “Beijing haze” topic in the first 9 days. The red line indicates the trend of the time series. The trend line is drawn with a moving average of a window size of 24

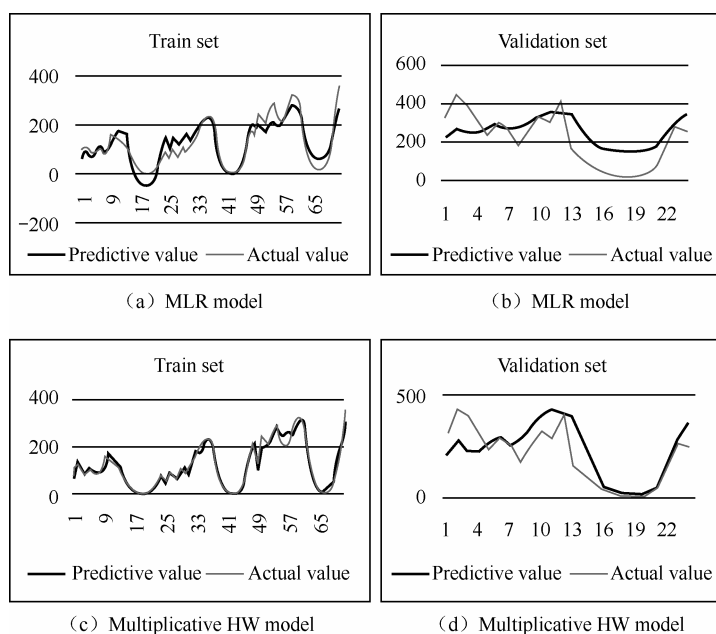


Figure 10-9 Results of “Beijing haze” popularity prediction from the MLR model and the multiplicative HW model

Take the “Christmas” topic as an example. As shown in Figure 10-10, the peak appears in December each year; the trend line is drawn with the move smoothing method using a window size of 12. The data from January 2008 to December 2009 were used as the train set, and the data from January to December of 2010 were used as the validation set. The results are shown in Figure 10-11.

3,000 samples were collected for the non-emergent topics, and 4,000 samples for the emergent topics. The historical data of both have a time span of 3 periods. The multiplicative HW model shows an average accuracy rate of 80.4% in predicting the trend of non-emergent topics, and an average accuracy rate of 84.7% in predicting the trend of emergent topics, while those of the MLR model are 63.7% and 71.4% respectively.

The HW model is much more accurate than the MLR. In addition, it also validates the seasonality of the hot topics is multiplicative season. The experimental results also show that the smoothing parameters usually remain stable in a certain range in both cases: $\alpha < 0.5$, $\beta < 0.5$, $\gamma < 0.07$ (non-emergent), $\alpha > 0.55$, $\beta > 0.5$, $\gamma < 0.07$ (emergent), which indicates that short-term emergent topics have more unstable level and trend terms, and more severe ups and downs, and therefore, they require frequent collection of the latest information; in other words, the dependence of its future popularity on historical data is weaker than

long-term type topics, while the historical data of long-term topics has stronger impact on the future popularity of these topics.

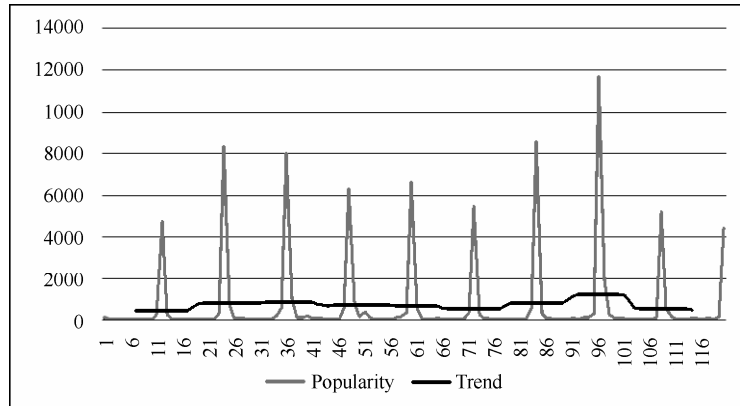


Figure 10-10 The grey line indicates the time series of the “Christmas” topic from January 2003 to December 2012. The black line indicates the trend of the time series. The trend line is drawn with a moving average of a window size of 24. The popularity reaches its peak in December each year

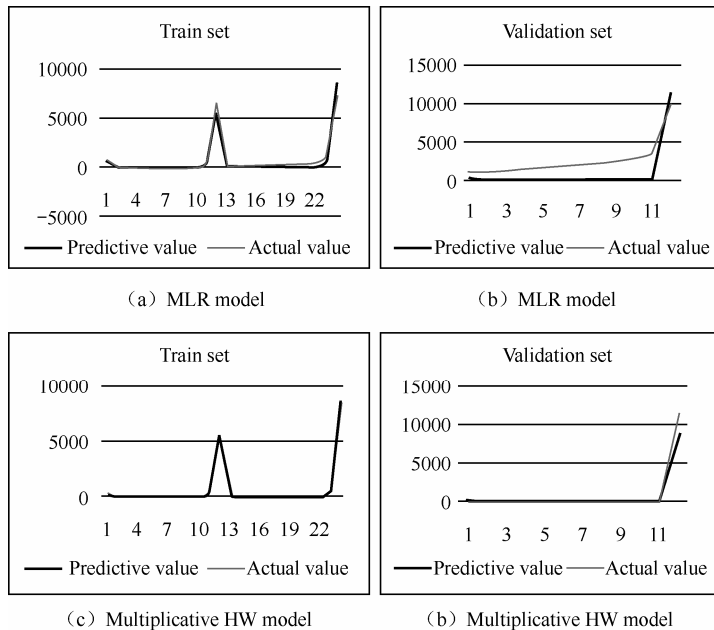


Figure 10-11 Results of “Christmas” popularity prediction from the MLR model and the multiplicative HW model

3. Minimum Historical Data

In order to achieve accurate predictions earlier, we carried out a research to find out the minimum periods of historical data required for making the most accurate prediction as far as the multiplicative HW model is concerned. For each topic, the last period of data was used to form the validation set. The results are shown in Figure 10-12. According to the results, the prediction for long-term non-emergent topics is the most accurate when three periods of historical data are collected, with an average accuracy rate of 0.813; the prediction for short-term emergent topics is the most accurate when two periods of historical data are collected, with an average accuracy rate of 0.858. This also re-verified the fact that the historical data of long-term topics has stronger impact on the future popularity, while the dependence of short-term non-emergent topics' popularity on historical data is weaker. When the number of historical data periods exceeds three, the accuracy cannot be significantly improved, because data in one period can reflect the seasonal information, data in two periods can reflect the trend information. As the levels and trends are in constant changes, earlier historical data cannot reflect the current popularity. Therefore, the accuracy reaches saturation after a time length of 3 periods.

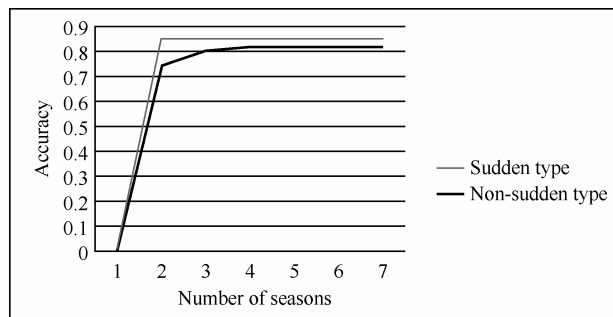


Figure 10-12 Average accuracy rates resulting from different season numbers used as train sets examined on two types of topics

4. Comparative Analysis

In order to validate the time series method, the SH Model based on the sample set is used for comparative analysis. Gabor Szabo and Bernardo A. Huberman found that the early popularity and late popularity exhibit linear relationship in logarithmic form: $\ln y = \ln x + b$, where x is the early popularity, a known data; y is the late popularity, to be predicted; b is a parameter trained from the sample set.

The multiplicative HW and SH models are used for comparison. 4,000 emergent-type hot topics serve as the sample set of the SH model; the trained parameter $b = 0.267$, and “Beijing haze” is the topic of prediction. Data in the first 7 days are used for early popularity, data in the 8th day are used for late popularity. The popularity on the 8th day predicted by the SH model is 9154, that predicted by the time series is 5488, and the actual popularity is 4913. The error rates of the two models as to the two topics are shown in Table 10-4.

Table 10-4 Average accuracies of SH and HW models in calculating the two topics

	Emergent	Non-Emergent
SH	29.5%	23.0%
HW	85.1%	80.4%

The reason why the SH model is not ideal in predicting the popularity of hot topics is that the SH model relies on the sample set, and the development law of each hot topic has its own characteristics. For example, each has different peak and base values, making it difficult to find the right sample set. However, time series method does not depend on the sample set, but on analyzing the historical data of the hot topic to be predicted.

It is easy to obtain historical data for a historical popularity based prediction model, which is suitable for predicting the long-term popularity of online contents. It can also be applied to predict the popularity of various online contents, because all posts, news, video, and microblogs have the historical data of popularity. However, predicting long-term popularity with premature historical data can lead to inaccurate predictions, while the use of late historical data can lead to belated predictions.

The advantage of a model based on the network structure is that the network structure factor is taken into consideration, which is more accurate than the prediction model based on the historical popularity, while its limitation is the same as that of the historical data based predication; that is, predicting long-term popularity using premature historical data and network structure information can lead to inaccurate predictions, while using late historical data and network structure information can lead to late predictions. Therefore, it is not suitable for social networks having vague features of network hierarchy structure, such as, video sharing websites and forum websites.

The predication model based on user behaviors takes into consideration the direct factors that enhance the popularity of online contents, while analysing user behaviors and

the status of online contents; therefore, the predictions are timely. However, this model assumes that all users have the same habits and behaviors, while individuals on social networks are actually different from one to another.

In contrast, a prediction model based on time series predicts the future popularity of certain online content by using the historical data of the contents themselves, which is more suitable for the online contents with more complex popularity evolution patterns. In general, the more popular the online content is, the more complex the popularity evolution pattern is. Of course, the time series method also has its drawbacks; that is, it is only suitable for short-term prediction, and is not accurate enough for predicting future trends farther beyond.

10.7 Information Source Location

10.7.1 Concept of Information Source Location

In the study of the diffusion process, how to determine the source of the diffusion based on the observed results of diffusion is a fundamental problem. Research results on this issue can usually be applied to spam management, computer virus prevention, rumors prevention and control on social networks and other areas.

Because of the conveniency and strong interactivity of social networks, information on social networks can spread very fast and widely on them, which also lead to the uncontrolled diffusion and spread of enormous false and illegal information. To identify the source of malicious information by means of the information source location technology and some effective methods is the key to control the diffusion of false and illegal information on social networks. The basic goal of information source location is to find out the initial source of information diffusion.

According to the existing researches on information source location, information source location is defined as follows: to determine the initial source node of information diffusion on a network in the condition of knowing the observed result, given the attributes of the underlying network structure, the mode of information diffusion, etc. Oftentimes, our observations of the results of diffusion are not complete, for we can only observe part of the overall result, which adds difficulty to the source location of the information. In addition, due to the diversity and uncertainty of the underlying network structure, and the different

characteristics of various diffusion modes themselves, the research of information source location technology faces many challenges.

The following part will introduce the main research results of information source location. In the model proposed by Vincenzo Fioriti et al.^[36], the dynamic importance of each node is calculated, in order to sort the nodes. Given the diffusion result - the undirected contact graph, it is possible to identify multiple source nodes or neighbors near these source nodes. This method behaves very well when the diffusion result exhibits a graph structure similar to a tree structure but is poorly behaved in other cases.

In their studies on identifying the starting point of a diffusion process in complex networks^[37], Cesar Henrique Comin et al. mainly analyzed the characteristics of the centrality of the source node and identified it by using an improved centrality measurement method. This method has a high success rate in ER networks and scale-free networks. The method was experimented in three different diffusion modes. The results show that the effect is best when the diffusion mode approximates the Snowball.

In the model^[38] proposed by Andrey Y. Lokhov et al., information diffusion is assumed to fit the SIR model. A reasoning algorithm based on the dynamic message transfer equation is used for source location, and with each node serving as the source node, the probabilities of the other nodes respectively in the three states of SIR are calculated. The algorithm is also effective in the case where only a part of the diffusion result is observed, but the complexity is relatively high in the case where the number of nodes is high in the network.

Nino Antulov-Fantulin et al. proposed a statistical reasoning framework based on Maximum Likelihood Estimation^[39] for source location. The study assumes that the diffusion process fits the SIR model and obtains a sorted list of possible nodes by likelihood estimation based on the observed diffusion result on any network. The study performs well on different network structures.

Each of the above methods has its own different applicable conditions. The basic ideas of these methods can be divided into two categories: one is the measurement (grading) of node attributes, and the other is source location based on statistical reasoning. In the following sections, we will first introduce two representative methods - the source location method based on centrality measurement, and the source location method based on a statistical reasoning framework, and then we will introduce a multi-source information source location method^[41] based on the reverse

diffusion and node partitioning targeted for the conditions of multi-source concurrency and incomplete observation.

10.7.2 Source Location Methods Based on Centrality

In study on identifying the starting point of the diffusion process in complex networks, Cesar Henrique Comin et al. analyzed the characteristics of the centrality of the source node and identified the source node by using an improved centrality measurement method. The study mainly examined two different networks: the ER network and the scale-free network, and both networks were experimented.

In reality, information diffusion exhibits different modes. For example, in a computer virus diffusion network, when a new virus appears, the node that is infected with the virus will diffuse it to all its neighbors; almost all of the neighbors will be infected, and the process keeps on; however, in the information diffusion process on social networks, there is a certain probability that an infected node infects its neighbors. Normally not all of them are infected, and only a part of the nodes are infected due to their interest in the information. Obviously, the performance of the information source location method will be affected by the different characteristics of different modes of diffusion. The study mainly considered the following three types of diffusion modes, and conducted experiments:

1. Snowball (Also Known As Dilation)

This diffusion method is similar to the classic breadth-first search algorithm, which corresponds to spam diffusion in the real world, i.e., information is sent from one node to all contacts.

2. Diffusion

The Random Walk algorithm can be referred to in this diffusion method. A node chooses one of its neighbor nodes for diffusion. Each neighbor node has a probability of being diffused.

3. Contact Process

This diffusion method can be viewed as a classic virus diffusion. Each node has a certain probability to infect its neighbors.

The source location method proposed in this study mainly deals with calculating the centrality of each node. In the past studies, the four major centrality measuring methods are “Degree”, “Closeness”, “Betweenness”, and “Eigenvector”.

The first measurement method, i.e., degree, uses the classical definition of degree given in the graph theory; that is, the number of edges associated with a node. d_{ij} represents the length of the shortest path between nodes i and j , so the average shortest distance passing node i l_i is:

$$l_i = \frac{1}{n-1} \sum_{j, j \neq i} d_{ij} \quad (10-30)$$

The second measurement method is closeness. In the following equation, the closeness of node i C_i is the reciprocal of the average shortest distance of node i :

$$C_i = \frac{1}{l_i} \quad (10-31)$$

We can see that in the “closeness” method, the distance between one node and the other nodes is used to measure the centrality of the node. Obviously, if the average distance between the node and the other nodes is small, it approximates the “center” of the network; in other words, it has higher centrality.

The third measurement method is “betweenness”. As shown in the following equation, the “betweenness” of node i is:

$$B_i = \sum_{s, t, s \neq t, s \neq i, t \neq i} \frac{n_{st}^i}{n_{st}} \quad (10-32)$$

Wherein, n_{st}^i is the number of the shortest paths between node s and node t via node i , and n_{st} is the number of shortest paths between node s and node t . It can be seen that the betweenness method measures the centrality of a node by testing whether the node is on the shortest path between other nodes. If a node is on the shortest path between many other nodes, it is more like a “hub” which has higher centrality.

The fourth measurement method is “eigenvector”. Eigenvector centrality follows the principle that when a node is connected to other high-level nodes, its importance becomes higher. Let s_i represent the score of the i -th node, and A denote the adjacency matrix of the network, the score obtained by the i -th node is the sum of the scores of all its neighbor nodes. Therefore,

$$s_i = \frac{1}{\lambda} \sum_{j=1}^N A_{ij} S_j \quad (10-33)$$

Wherein, λ is a constant. The above equation can be rewritten as:

$$As = \lambda s \quad (10-34)$$

The eigenvector of the largest eigenvalue obtained by this equation represents the eigenvector of the node.

In this study, the diffusion process on the network can be simulated in the following way: given that some seed nodes are assumed as the starting nodes, the underlying network is sampled by the algorithm corresponding to the three different types of diffusion mentioned earlier, to get a subgraph.

Cesar Henrique Comin et al. measured the centrality of the nodes in the subgraph achieved after the sampling on the ER network and the scale-free network, and found that the degree of the nodes was almost unchanged after sampling because of the local variables. Therefore, the deviation caused by sampling can be eliminated when the measured centrality value is divided by the degree of the node. The unbiased betweenness is defined as follows:

$$\hat{B}_i = \frac{B_i}{(k_i)^r} \quad (10-35)$$

Wherein, B_i is the biased betweenness, k_i is the degree, and an appropriate empirical value selected for r by experiment is $r = 0.85$. The result of measurement achieved by using the improved method shows that the centrality of the source node is significantly higher than that of other nodes, and the source node can be well identified on the ER network and the scale-free network. In addition, Cesar Henrique Comin et al. experimented in three different modes of diffusion and found that the effect of this source location method is best when the diffusion mode approximates the Snowball.

Example 10-5 Example of centrality based source location calculation

Calculate the relevant properties of node a in Figure 10-13.

Node a has a degree of 2. As this is a directed graph, both the in-degree and the out-degree are 2.

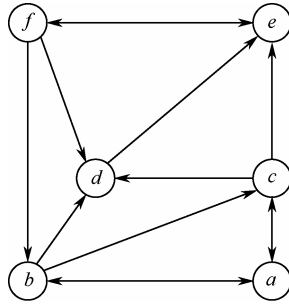


Figure 10-13 Example of centrality source location calculation

The shortest distances from node a to node b , node c , node d , node e and node f are respectively 1, 1, 2, 2, 3, so the average shortest distance to node a is $l = \frac{1+1+2+2+3}{5} = 1.8$.

The closeness of node a is $C = 1/1.8$.

The betweenness of node a is $B = \frac{1}{1} = 1$, because there is only one shortest path between node c and node b , and only the shortest path between node c and node b passes through node a .

10.7.3 Source Location Methods Based on Statistical Reasoning Framework

Nino Antulov-Fantulin et al. proposed a statistical reasoning framework based on Maximum Likelihood Estimation, which is based on the observed diffusion process on any network.

In this study, the SIR model is used as the diffusion model. In the network G , the nodes have three states: susceptible (S), infected (I) and recovered (R). The diffusion process is simulated using discrete time steps. The probability that a susceptible node is converted to the infected state is p , and the probability that the infected node is converted to the recovery state is q . Let θ be the original infection node, assuming that the time step experienced by the diffusion process is known and used as a parameter for estimating the source node by reasoning.

Based on the above assumptions, the source location problem is defined as follows:

The random vector $\vec{R} = (R(1), R(2), \dots, R(N))$ represents the infection of the node

before a certain time threshold T . The random variable $R(i)$ is a Bernoulli random variable, and if the node is infected before the time point T , the corresponding value is 1; otherwise, the value is 0.

Suppose we have observed the SIR model diffusion result of a known (p, q) and T , and the set of all nodes is $S = \{\theta_1, \theta_2, \dots, \theta_N\}$, with a limited number of nodes. We get the following problem of maximum likelihood:

$$\hat{\Theta} = \arg \max_{\Theta \in S} P(\Theta | \vec{R} = \vec{r}) \quad (10-36)$$

Wherein, $\Theta \in S$ is the all possible sources of diffusion. According to Bayes theorem, we can see that:

$$\hat{\Theta} = \arg \max_{\Theta \in S} P(\vec{R} = \vec{r} | \Theta) \quad (10-37)$$

Algorithm 10-3 represents a process that uses a maximum likelihood estimation algorithm to perform calculations. The main idea of maximum likelihood estimation is, with the experimental result being already known, to find the experimental condition that is most favorable (i.e., of the largest likelihood) for getting the experimental result through the algorithm. The experimental result here means the observed diffusion result that is already known. At the same time, some conditions, including the parameters (p, q) and T of the SIR model, are also known. The unknown experimental condition is the source of diffusion.

Algorithm 10-3 Maximum Likelihood Based Source Node Estimation Algorithm:

$(G, p, q, \vec{r}_*, T, S, n)$

Input: network structure G , SIR process parameter (p, q) , possible source node set $S = \{\theta_1, \theta_2, \dots, \theta_N\}$, observed dissemination result \vec{r} , cutoff time threshold T , number of simulations n

for each $\theta_j \in S$ (a prior set of possible source nodes) **do**

Likelihood estimation function calling $(G, p, q, \vec{r}_*, T, n)$

Save $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta_j)$

end for

Output 1: θ_k and maximum likelihood estimation $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta_k)$

Output 2: source nodes sorted based on likelihood $\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta_k)$ in set $S = \{\theta_1, \theta_2, \dots, \theta_N\}$

The n in the parameters is the number of simulations run for a set of source node

candidates.

The similarity φ of two different diffusion results is judged in two ways (XNOR and Jaccard), which are denoted as $\overline{\text{XNOR}}(\vec{r}_1, \vec{r}_2)$ and Jaccard (\vec{r}_1, \vec{r}_2) , respectively. Later, Nino Antulov-Fantulin et al. defined three different likelihood estimation functions: AUCDF, AvgTopK, and naive Bayesian. The first two uses the similarity calculation method mentioned earlier, and the naive Bayesian method uses its own similarity calculation method. Although the three algorithms are different, they share the same main idea; that is, to calculate the likelihood, i.e., the probability of achieving the experimental results under different experimental conditions. In the following part, we will introduce these three algorithms.

Algorithm 10-4 represents the algorithm for the AUCDF estimation function.

Algorithm 10-4 AUCDF Estimation Function $(G, p, q, \vec{r}_*, T, S, n)$
Input: network structure G , SIR process parameter (p, q) , observed diffusion result \vec{r}_* , calculated source node θ , cutoff time threshold T , number of simulations n for $i = 1$ to n (number of simulations) do Run SIR simulation (p, q) , wherein $\Theta = \theta$, get the propagation result $\vec{R}_{\theta, j}$, stop when the time threshold T is reached Calculate and save $\varphi(\vec{r}_*, \vec{R}_{\theta, j})$ end for Calculate the actual distribution function: $\hat{P}(\varphi(\vec{r}_*, \vec{R}_{\theta, j}) \leq x) = \frac{\sum_{i=1}^n 1_{[0, x]}(\varphi(\vec{r}_*, \vec{R}_{\theta, j}))}{n}$ Calculate the likelihood: $\text{AUCDF}_{\theta} \int_0^1 \hat{P}(\varphi(\vec{r}_*, \vec{R}_{\theta, j}) \leq x) dx$ Output: $\hat{P}(\vec{R} = \vec{r}_* \Theta = \theta) = 1 - \text{AUCDF}_{\theta}$

Algorithm 10-5 represents the algorithm for the AvgTopK likelihood estimation function

Algorithm 10-5 AvgTopK Likelihood Estimation Function $(G, p, q, \vec{r}_*, T, \theta, n)$
Input: network structure G , SIR process parameters (p, q) , observed diffusion result

\vec{r}_* , calculated source node θ , cutoff time threshold T , number of simulations n

for $i = 1$ to n (number of simulations) **do**

Run SIR simulation (p, q) , wherein $\Theta = \theta$, get the propagation result $\vec{R}_{\theta,J}$, stop when the time threshold T is reached

Calculate and save $\phi(\vec{r}_*, \vec{R}_{\theta,J})$

end for

Sort the ratings $\{\phi(\vec{r}_*, \vec{R}_{\theta,J})\}$ in descending order;

Averaged of k maximum ratings:

$$\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta) = \frac{1}{k} \sum_{i=1}^k \{\phi(\vec{r}_*, \vec{R}_{\theta,J})\}_{\text{sorted}}$$

Output:

Likelihood $\bar{P}(\vec{R} = \vec{r}_* | \Theta = \theta)$

Algorithm 10-6 represents the algorithm for the Naive Bayesian likelihood estimation function.

Algorithm 10-6 Naive Bayesian Likelihood Estimation Function $(G, p, q, \vec{r}_*, T, \theta, n)$

Input: network structure G , SIR process parameters (p, q) , observed diffusion result \vec{r}_* , calculated source node θ , cutoff time threshold T , number of simulations n

Wherein $m_k = 0 : \forall k \in V$ from G ;

for $i = 1$ to n (number of simulations) **do**

Run SIR simulation (p, q) , where $\Theta = \theta$, get the propagation result $\vec{R}_{\theta,J}$, stop when the time threshold T is reached

Update $m_k = m_k + 1$ for k being infected in $\vec{R}_{\theta,J}$

end for

Calculate:

$$\hat{P}(\vec{r}_*(k) = 1 | \Theta = \theta) = \frac{m_k + \epsilon}{n + \epsilon}, \forall k \in G$$

Calculate log likelihood:

$$\begin{aligned}
& \log(\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta)) \\
&= \sum_{\{k: \vec{r}_*(k)=1\}} \log(\hat{P}(\vec{r}_*(k)=1 | \Theta = \theta)) \\
&+ \sum_{\{j: \vec{r}_*(j)=0\}} \log(1 - \hat{P}(\vec{r}_*(j)=1 | \Theta = \theta))
\end{aligned}$$

Output:

Likelihood $\log(\hat{P}(\vec{R} = \vec{r}_* | \Theta = \theta))$

Nino Antulov-Fantulin et al. tested the performance of the likelihood estimation algorithm on different network architectures. After the list of potential source nodes output by the algorithm in node set S is achieved, the ranking of the actual source node in this output list is checked. Experiments show that this method performs well in various network environments.

10.7.4 Multiple Information Source Location Methods

Oftentimes, we can only observe some of the diffusion results. In a diffusion mode that fits the SIR model, part of the nodes will shift from the infected state into the recovery state, making it more difficult for information source location. In addition, the diffusion results derived from multiple-source diffusion make it harder for us to determine the true source of information. This section introduces a method of multi-point source location based on sparsely infected nodes, which solves the problem of incomplete observations and multiple-source diffusion. The method consists of the following three steps: First, detect the recovered nodes in the network using a reverse diffusion method. Secondly, partition all infected nodes by using a partitioning algorithm, thus to turn the issue of multi-point source location into the issue of multiple independent single-point source location. Finally, determine the most likely source node in each partition.

1. Step 1: Reverse Dissemination Method

In real life, it is very difficult to locate the source of rumors only by observing several infected nodes. Considering the recovery of nodes from the infected state, in order to supplement such lost information and simplify the problem of source node location, we proposed a score-based reverse diffusion method to find the recovered nodes and infected node set. Algorithm 10-7 shows an algorithm for the scoring function based on the reverse

diffusion method.

Algorithm 10-7 Scoring Function Based on The Reverse Diffusion Method

Input: a social network G , node set V and connection set E contained in the network, and part of the observed infected node set $I \in V$; a constant basesore.

Output: an extended infected node set $I^* \in V$, the infected node set which contains all the recovered nodes, the infected node set, both observed and unobserved, and the nodes that are associated with the infected nodes but are not infected. In addition $I \in I^* \in V$

Initialize the nodes, assign the unique label C and the unique score $S_c, \forall n \in V$:

$$C_n, S_{c_n} = \begin{cases} 1, & n \in I \\ 0, & \text{otherwise} \end{cases}$$

Initialize the node set $I' = I$

for iter 1 to N_{step} **do**

for $n \in I'$ **do**

for $i \in n_{\text{neighbors}}$ **do**

update $I' = I' \cup i, C_i = 1$

update $S_{C_i} = S_{C_i} + S_{C_n}$, if $C_i = 0$

for $n \in V$ **do**

if $S_{C_n} > \text{basescore}, I^* = I^* \cup n$

Return extended infected node set I^*

A new network is created for location analysis by implementing Algorithm 10-7. This network is called the expanded infection network.

2. Step 2: Infected Node Partitioning

Use the “divide and rule” approach to change the multi-point source location into an issue of multiple independent single-point source location. In the following part, we will briefly introduce three methods of partitioning:

1) Method based on modularity

The result of a good partitioning method should exhibit relatively dense connections between nodes in the same partition, but relatively sparse connections between nodes in different partitions. This can be measured by the modularity function. Modularity refers to the difference between the proportion of the edges connecting the internal vertices of the

community structure in a network and the expected value of the proportion of the edges connecting the internal vertices of the community structure in another random network.

The greater the value of this function, the better the effect of partition. This method can be rewritten in the form of a matrix, and the modularity is represented by the eigenvector of the matrix.

2) Method based on edge betweenness

Linton C Freeman first proposed to measure the centrality and influence of nodes in the network by using betweenness. Edge betweenness is defined as the number of the shortest paths passing an edge in the network. If an edge has high edge betweenness, it means that it is the edge that bridges the two partitions of the network. The partitioning steps based on the edge betweenness method are as follows:

- (1) Calculate the edge betweenness of all edges in the network;
- (2) Remove the edge with the highest edge betweenness;
- (3) Recalculate the edge betweenness of all edges;
- (4) Repeat step (2) until all edges are gone.

3) MMSB (Mixed Membership Stochastic Model)

This method is based on the assumption that nodes infected by the same source node are more likely to have connections, and that nodes infected by different source nodes rarely generate connections. Given a graph $G = (N, Y)$ which contains N nodes and the connection set $Y(Y(p, q) \in \{0, 1\})$. K is the number of potential partitions. The goal of MMSB is to obtain the parameters α and β by calculating the maximum likelihood of the edges.

$$P(Y | \vec{\alpha}, \beta) = \int_{\pi} \sum_{Z_S} (\prod_{p,q} P(Y(p, q) | \vec{z}_{p \rightarrow q}, \vec{z}_{q \rightarrow p}, B) P(\vec{z}_{p \rightarrow q} | \vec{\pi}_p) P(\vec{z}_{q \rightarrow p} | \vec{\pi}_q) \prod_p P(\pi_p | \vec{\alpha})) d\vec{\pi} \quad (10-38)$$

$B(g, h)$ in $B_{K \times K}$ represents the probability that there is an edge between a node in partition g group and a node in partition h , and $\Pi = \pi_{N \times K}$ denotes the matrix formed by the probabilities of nodes in each partition.

3. Step 3: Node Set Source Location

Through the above two steps, the multi-point source location issue is transformed into an issue of multiple independent single-point source location. Four measures related to centrality - Degree, Closeness, Betweenness, and Eigenvector, are adopted for evaluation. In this part, the method used by Cesar Henrique Comin et al. mentioned above in the

single-point source location is adopted, which we will not go into details here.

Experiment on three different types of artificial networks: random regular network, BA network, and ER network. All networks generate 5000 nodes through NetworkX. The probability of infection in the SIR model is 0.9 and the recovery probability is 0.2. Test the effect of different community detection algorithms on infected node partition. The experimental results show that the method based on the main features is superior to other community partitioning algorithms. Thus, the method based on the main feature is selected to evaluate the infected node in subsequent experiments.

Test the proposed multi-point source location solution on different networks. The three networks are: the random regular network with 5000 nodes, where the degree of each node is 3; ER network with 5000 nodes, where the probability of edge generation is 0.01; BA network with 5000 nodes, where each new node generates two edges to reach the existing nodes. The results of the experiment are shown in Figures 10-14 and 10-15. The solution achieved good results on the random regular network, which could effectively discover the source nodes.

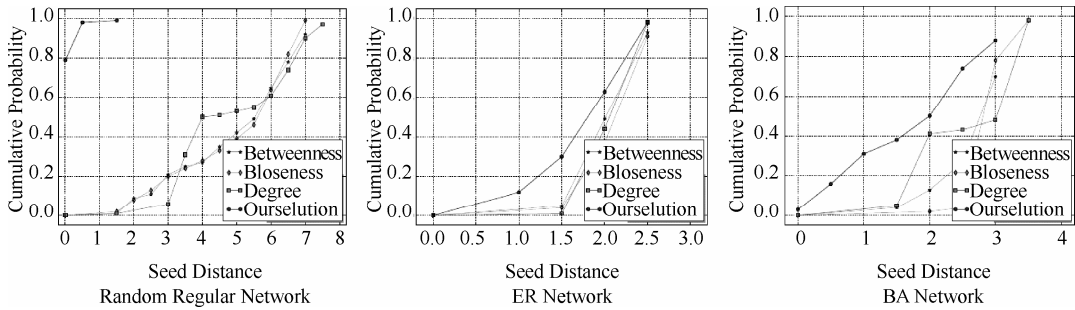


Figure 10-14 The cumulative probability distribution of the average distance between the real source node and the calculated source node, the number of source nodes $k=2^{[41]}$

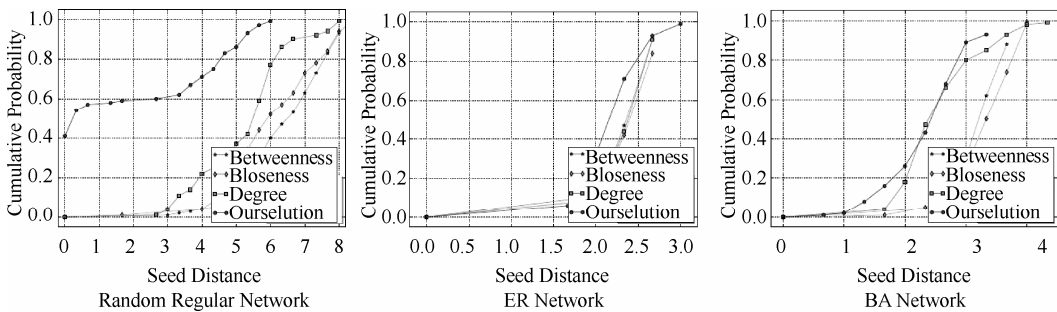


Figure 10-15 The cumulative probability distribution of the average distance between the real source node and the calculated source node, the number of source nodes $k=3^{[41]}$

10.8 Summary

The study of information diffusion in social networks has become one of the most challenging and promising fields of research. This chapter introduces a number of mainstream models of information diffusion in social networks and explores the methods for predicting the trends of information diffusion and for tracing the sources of information.

Despite the numerous studies and certain achievements in the field of information diffusion in social networks have been made, there are still many problems that require further exploration.

(1) Model validation and evaluation methods: the existing model validation methods are mainly based on random data for verification, or computer simulated data for analysis. However, for a more scientific approach, a unified standard test set should be established by screening the typical examples of diffusion, to assess the advantages and disadvantages of different diffusion models, and define the scope of applications of the algorithms.

(2) The intrinsic rule of multi-factor coupling: most of the existing references deal with information diffusion from a single perspective, such as, the topology of the network where the information is diffused, the rules of individual interactions, and so on. However, the diffusion of information in reality is the typical evolution process of a complex system, which requires a comprehensive consideration of multiple factors including individual interaction, network structure and information characteristics, in order to describe information diffusion in online social networks in a more accurate manner.

(3) Dynamic changes in social networks: most of the existing methods of information diffusion analysis are based on the static network topology; however, in the real social networks, the network of relationship between users changes over time. It is necessary to add the attribute of dynamic change into the information diffusion model. In addition, the existing algorithms are mostly based on serial or time step models. Large-scale parallel distributed algorithms are needed to improve the efficiency of processing.

It is extremely important and challenging to study the mechanism of information diffusion in social networks and to understand the law of information diffusion. There are still a number of important issues crying out for solutions. We expect more exciting results

in the future.

References

- [1] Mark S. Granovetter. The strength of weak ties[J]. American journal of sociology, 1973:1360-1380.
- [2] Stratis Ioannidis, Augustin Chaintreau. On the strength of weak ties in mobile social networks[C]. Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. ACM, 2009: 19-25.
- [3] Stephan Ten Kate, Sophie Haverkamp, Fariha Mahmood, Frans Feldberg. Social network influences on technology acceptance: A matter of tie strength, centrality and density[J]. BLED 2010 Proceedings, 2010, 40.
- [4] Paul S. Adler, Seok-Woo Kwon. Social capital: Prospects for a new concept[J]. Academy of management review, 2002, 27(1): 17-40.
- [5] Jichang Zhao, Junjie Wu, Xu Feng, Hui Xiong, Ke Xu. Information propagation in online social networks: a tie-strength perspective[J]. Knowledge and information systems, 2012, 32(3): 589-608.
- [6] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, Lada Adamic. The role of social networks in information diffusion [C]. Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 519-528.
- [7] John Scott. Social network analysis: developments, advances, and prospects[J]. Social network analysis and mining, 2011, 1(1): 21-26.
- [8] Mor Naaman, Jeffrey Boase, Chih-Hui Lai. Is it really about me?: message content in social awareness streams[C]. Proceedings of the 2010 ACM conference on Computer supported cooperative work. ACM, 2010: 189-192.
- [9] Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng. Why we twitter: understanding microblogging usage and communities[C]. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007: 56-65.
- [10] Mike Thelwall, Kevan Buckley, Georgios Paltoglou. Sentiment in Twitter events[J]. Journal of the American Society for Information Science and Technology, 2011, 62(2): 406-418.
- [11] Seth Myers, Jure Leskovec. Clash of the Contagions: Cooperation and Competition in Information Diffusion[C]. ICDM. 2012, 12: 539-548.
- [12] Mark Granovetter. Threshold models of collective behavior[J]. American journal of sociology, 1978: 1420-1443.
- [13] Jacob Goldenberg, Barak Libai, Eitan Muller. Talk of the network: A complex systems look at the

- underlying process of word-of-mouth[J]. *Marketing letters*, 2001, 12(3): 211-223.
- [14] Jacob Goldenberg, Barak Libai, Eitan Muller. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata[J]. *Academy of Marketing Science Review*, 2001, 9(3): 1-18.
- [15] Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, Andrew Tomkins. Information diffusion through blogspace[C]. *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004: 491-501.
- [16] Xiaodan Song, Yun Chi, Koji Hino, Belle L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking[C]. *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007: 191-200.
- [17] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, Hiroshi Motoda. Behavioral analyses of information diffusion models by observed data of social network[M]. *Advances in Social Computing*. Springer, 2010: 149-158.
- [18] Kazumi Saito, Masahiro Kimura, kouzou Ohara, Hiroshi Motoda. Selecting information diffusion models over social networks for behavioral analysis[M]. *Machine Learning and Knowledge Discovery in Databases*. Springer, 2010: 180-195.
- [19] Luke Dickens, Ian Molloy, Jorge Lobo, Paul-Chen Cheng, Alessandra Russo. Learning stochastic models of information flow[C]. *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, 2012: 570-581.
- [20] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, Wolfgang Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs[C]. *Proceedings of the 3rd conference on Online social networks*. USENIX Association, 2010: 3.
- [21] Adrien Guille, Hakim Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks[C]. *Proceedings of the 21st international conference companion on World Wide Web*. ACM, 2012: 1145-1152.
- [22] William O. Kermack, Anderson G. McKendrick. Contributions to the mathematical theory of epidemics[J]. In *Proceedings of the Royal Society of London*, 1927, 115(772): 700-721.
- [23] William O. Kermack, Anderson G. McKendrick. Contributions to the mathematical theory of epidemics. II. The problem of endemicity[J]. *Proceedings of the Royal society of London. Series A*, 1932, 138(834): 55-83.
- [24] Michelle Girvan, Mark EJ Newman. Community structure in social and biological networks[J]. *Proceedings of the National Academy of Sciences*, 2002, 99(12): 7821-7826.
- [25] Saeed Abdullah, Xindong Wu. An epidemic model for news spreading on twitter[C]. *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*. IEEE, 2011:

- 163-169.
- [26] Fei Xiong, Yun Liu, Zhen-jiang Zhang, Jiang Zhu, Ying Zhang. An information diffusion model based on retweeting mechanism for online social media[J]. *Physics Letters A*, 2012, 376(30): 2103-2108.
 - [27] Dechun Liu, Xi Chen. Rumor Propagation in Online Social Networks Like Twitter-A Simulation Study[C]. *Multimedia Information Networking and Security (MINES)*, 2011 Third International Conference on. IEEE, 2011: 278-282.
 - [28] Jaewon Yang, Jure Leskovec. Modeling information diffusion in implicit networks [C]. *Data Mining (ICDM)*, 2010 IEEE 10th International Conference on. IEEE, 2010: 599-608.
 - [29] Thomas F. Coleman, Yuying Li. A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables[J]. *SIAM Journal on Optimization*, 1996, 6(4): 1040-1058.
 - [30] Seth Myers, Chenguang Zhu, Jure Leskovec. Information diffusion and external influence in networks[C]. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012: 33-41.
 - [31] Alex Beutel, B. Aditya Prakash, Roni Rosenfeld, Christos Faloutsos. Interacting viruses in networks: can both survive? [C]. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012: 426-434.
 - [32] Gabor Szabo, Bernardo A. Huberman. Predicting the popularity of online content[J]. *Communications of the ACM*, 2010, 53(8):80-88.
 - [33] Peng Bao, Hua-Wei Shen, Junming Huang, et al. Popularity prediction in microblogging network: a case study on sina weibo [C]. *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013: 177-178.
 - [34] Kristina Lerman, Tad Hogg. Using a model of social dynamics to predict popularity of news[C]. *Proceedings of the 19th international conference on World wide web*. ACM, 2010: 621-630.
 - [35] Changjun Hu, Ying Hu. Predicting the Popularity of Hot Topics Based on Time Series Models [C]. *APWEB*, 2014.
 - [36] Vincenzo Fioriti, Marta Chinnici. Predicting the sources of an outbreak with a spectral technique[J]. *arXiv preprint arXiv:1211.2333*, 2012.
 - [37] Cesar Henrique Comin, Luciano Da Fontoura Costa. Identifying the starting point of a spreading process in complex networks[J]. *Physical Review E*, 2011, 84(5): 56105.
 - [38] Andrey Y. Lokhov, Marc M E Zard, Hiroki Ohta, et al. Inferring the origin of an epidemic with dynamic message-passing algorithm[J]. *arXiv preprint arXiv:1303.5315*, 2013.

- [39] Nino Antulov-Fantulin, AlenLancic, Hrvoje Stefancic, et al. Statistical inference framework for source detection of contagion processes on arbitrary network structures[J]. arXiv preprint arXiv:1304.0018, 2013.
- [40] B. Aditya Prakash, Jilles Vreeken, Christos Faloutsos. Spotting culprits in epidemics: How many and which ones? 2012[C]. ICDM. 2012, 12: 11-20.
- [41] Wenyu Zang, Peng Zhang, Chuan Zhou. Discovering Multiple Diffusion Source Nodes in Social Networks [C]. ICCS 2014.

Topic Discovery and Evolution

11.1 Introduction

With the rapid development of information technologies and the wide spread of information applications, online social networks are gradually replacing the traditional Internet services and have become a more convenient and efficient channel for information dissemination and access. In social networks, news and events happened at every second are reported timely in different regions and with diversified languages and moreover, they are spreading in social networks across geographical boundaries. Social networks feature rich information and complex contents, which is filled with topics that a large number of users may find interest on. How to extract and recommend topics of users' interest from the massive, dynamic and multi-source social network data, track topic development and dig into the development trend of events are very crucial for information decision in such a rapidly changing Internet age.

In the most research on topic discovery and evolution, a topic refers to an event or an activity that draws public attention, and all related events and activities^[1], where an event or an activity happens at a particular time and a particular place^①.

In today's booming social networks, as a data-intensive application, topic discovery and evolution in social networks has the following data characteristics:

① In some research on topic discovery and evolution, they focus on the discovery and evolution of a single event. However, from the point of algorithm there is no distinguished difference from the most works on topics. Hence, in this book, we don't claim this difference particularly between event discovery and topic discovery.

(1) The original occurrence location of topics in social networks is distributed. Users of social networks can launch topics at any time in any place and the corresponding locations are scattered. Basically, we can't predict the exact time or place of the occurrence of a topic.

(2) Topics in social networks transmit rapidly and with a large range. Through the globally-connected social networks, a topic can have a huge impact in the world within a few hours after its launch.

(3) Topics in social networks have a wide variety and they almost cover and contain everything, which results in complex feature of corresponding text and needs support from multi-domain feature knowledge. Therefore, traditional text analysis methods are not applicable to text analysis related to topics from social networks.

(4) The topic-related data from social networks are multi-sources. The initiator of a topic is usually not one person, which makes the structure of the topic complicated and causes contradictions and conflicts among opinions of the topic.

(5) The topic-related data from social networks are massive and relatively concentrated. In large social networks, the data generated every day is massive. For example, Facebook usually handles about 2.5 billion messages on average every day. But all these vast amounts of data are relatively concentrated in a few large social networks. On Chinese networks, QQ, Sina Weibo and WeChat include almost all topics of social network data.

(6) The topic-related data from social networks are dynamically and constantly updated. Because of the interaction of users in social networks, a user's attitude about a particular topic may change with that of the surrounding friends.

Based on the above data features, the methods of topic discovery and evolution in social networks, unlike traditional media monitoring, should be able to automatically discover and track the evolution of the topic with no need for too much human interventions. Besides, because the topic data of social networks are multi-sources, dynamic and massive, data discovering and tracking by manpower in social networks are almost impossible. So it is necessary to propose algorithms for computer programs on topic discovery and evolution in social networks to ensure automatic topic detecting and tracking by computer programs.

As a relatively novel research subject, topic discovery and evolution in social networks, totally different from that of traditional media, has not been studied and explored in-depth before and thus research on this issue still remains in a relatively preliminary stage.

This content of this chapter is organized as follows: in section 11.2, as one of the

theoretical base of topic-related research in social networks, the models and algorithms of topic discovery are introduced in details, including topic model based topic discovery, vector space model based topic discovery and term relationship graph based topic discovery. In section 11.3, the models and algorithms of topic evolution, as another theoretical base of topic-related research in social networks, are introduced, including the simple topic evolution, topic model based topic evaluation and adjacent time slice relation based topic evaluation. Section 11.4 is a brief summary of this chapter.

11.2 Models and Algorithms of Topic Discovery

Traditional topic detection is generally backgrounded by news story or scientific corpus, such as the famous TDT (Topic Detection and Tracking) project^[1], which regards discovering and tracking topics from news story as its goal. But the difference between news corpus or scientific corpus and current social network corpus is huge, so as for social network data, direct application of traditional methods may not lead to good results. Therefore, we must propose new methods or improve the traditional methods to adapt to the characteristics of social network data.

As an emerging social media, Twitter, along with the development of Internet, has attracted a large number of users. Meanwhile, the data published by twitter users have also grown geometrically. We take the following tweets generated by twitter as an example to analyze the data characteristics on the research of topic discovery in social networks.

(1) The scale of the data is large and the updating speed is fast, so as for algorithm on twitter data processing, the application efficiency under big data environments should be considered and online dynamic requirements should be met.

(2) The data content is brief. A tweet is usually limited to 140 characters, hence the messages on twitter are relatively brief with some even only containing one or two words.

(3) The noise of data is too much. Tweets posted by twitter users tend to be more casual, whose main content usually consists of personal issues and personal viewpoints, and due to the length limit of the content, it is often mixed with mispronounced characters, new words, network slang, abbreviations, emoticons, special labels [such as twitter in the # hashtag # (twitter tag)] etc. Of course, there are also stop words as in traditional text. These features are all noise information if they are handled with traditional methods. In short, twitter data contains little text information, usually of low quality^[4]. Consequently, this provides a great challenge to text processing and topic discovery.

With regard to features of twitter data, we can propose solutions to the problems of topic discovery targetedly. For example, as to large-scale data, we can use a distributed algorithm; as for online demands, we can apply an online machine learning algorithm; and about data briefness, we can adopt aggregation strategy. All in all, we should pay special attention to the differences between data in social networks and traditional data on topic discovery, only in this way can it guide us to put forward reasonable and effective solutions to data in social networks.

11.2.1 Topic Model Based Topic Discovery

Conceptually speaking, topic model refers to a statistical model applied to find abstract topic in a series of documents.

The topic model is derived from the Latent Semantic Analysis (LSA) model^[7], which was proposed by Scott Deerwester in 1990. LSA model mainly adopts Singular Value Decomposition (SVD) method. Although probabilistic is not introduced in this method, it does provide a foundation for the later development of the topic model. In brief, the probabilistic LSA is actually the Probabilistic Latent Semantic Analysis (PLSA) model^[12], and afterwards, a more perfect Latent Dirichlet Allocation (LDA) model^[5] is proposed with further probabilistic parameters, which forms a hierarchical Bayesian graph model. With the gradual development of research progress, the whole theory of the topic model tends to improve gradually, accordingly, the topic model has been widely applied in various fields. Topic model is now not only used in text processing, also adopted in bioinformatics and image processing, which achieves good results in all of these areas. Among all the topic models, LDA model has been the most widely used one for it has such advantages as solid statistical base and flexibility to adapt to different task requirements.

1. LDA Model Introduction

Topic model bases on such assumption: it is based on the bag-of-words model, namely, in this model words in the document feature interchangeability, i.e., each word is independent of others, and the exchanging order has no effect on the document. Of course, upon this assumption, the true natural language has been simplified so as to facilitate computer processing.

After we have finished an article, there will be a term distribution^①, i.e., the proportion of each term in this document, here represented with $p(w|d)$ (because a topic model involves so many symbols, we listed all these symbols in Table 11-1). It is easy to calculate the formula $p(w_i|d) = N_i / N$, where N_i represents the occurrence number of the word w_i in the document, and N represents the total number of words in the document.

According to this idea, if $p(w|d)$ of a document is identified, we can quickly generate a document based on the probability distribution.

Table 11-1 Involved symbols of a topic model

Symbols	Implication	Symbols	Implication
N	Document length (number of words)	M	Document number
D	Document	z	Topic
W	Terms	α	Dirichlet distribution parameter, determining the distribution of document topic
β	Dirichlet distribution parameter, determining the distribution of topic terms	E	Distribution parameter, determining the distribution of the document tags
K	Topic number	θ	Topic distribution in the document
ϕ	Topic term distribution	U	Users
θ^μ	Topic distribution of user μ	ϕ^z	Term distribution of topic z
ϕ^B	Term distribution of background B	Π	Binomial distribution, for select control
Λ	Topic distribution associated with tags	Z	Topic set
S	Term set associated with a topic	TF	Term frequency
DF	Document frequency	E	Edge formed by co-occurrence terms

But obviously this process is not in line with our usual writing process. When we write an article, generally, we first select a topic, and then select a number of words related to this topic to enrich the article, in this way an article is generated. Therefore, according to this approach, the completion of an article is divided into two steps: firstly, selecting a topic of a document; and secondly, selecting words related to the selected topic, repeating this process in turn and generating one word at a time until the number of words specified in the

① The difference between word and term: in a document represented by a term space, every term is a dimension, and the number of terms in a document refers to the number of different words. However, during the number counting for words, two same words can be counted repeatedly.

document is reached.

Topic model is a statistical modeling of the above-mentioned process, where the differences remain in that there is more than one topic in the assumption document of a topic model. According to the generating process of a topic model document, a brief description is as follows:

Assume that we already know the topic distribution $p(z|d)$ of a document and the term distribution $p(w|z)$ of a topic.

When generating a word in a document, we first select a topic z according to the topic distribution, and then select a term w according to the term distribution under the topic. Hence, the word generation process in a document can be formulated from the probability perspective and simply interpreted as

$$p(w|d) = \sum_z p(z)p(w|z)p(z|d) = p(d) \sum_z p(z|d)p(w|z) \quad (11-1)$$

Here $p(w|d)$ is known, and $p(w|z)$ and $p(z|d)$ is unknown; assuming that there are M documents, and the length of each document d is N , i.e., there are N words, and there are K optional topics, then $p(w|d)$ is a vector of $M \times N$, $p(z|d)$ is a vector of $M \times K$, and $p(w|z)$ is a matrix of $K \times N$.

From Eq. (11-1), we can observe that this process inserts an intermediate layer more directly by the term distribution generated by words of the document -- topic layer, which is invisible in an actual document. Therefore, in some reference this layer is called Latent Structures of a document, or by intuitive understanding this is what we want to express--topic.

Here is a simple example (where although the data is artificially constructed, it is consistent with our understanding on real data).

Suppose that we have a document set $D = \{\text{Document 1, Document 2, Document 3}\}$, a term list $V = \{\text{movie, music, tax, government, student, teacher, amount, art, principal}\}$, and a topic set $Z = \{\text{art, budget, education}\}$, the optional words in each document are from V (actual term data is undoubtedly much larger than the set V here, and examples here are only for description convenience), and optional topics in each document are from the topic set Z . Here we artificially construct the proportion of words of each document but in practice it can be obtained by taking the ratio of the frequency of a specific word and the total number of words.

In this example, the matrix $p(w|d)$ constructed by documents and terms is as below, where the horizontal represents different documents, while the vertical represents different

terms and the corresponding number is the probability of a term in a document.

	movie	music	tax	government	student	teacher	money	art	principal
Document 1	0.27	0.19	0	0.027	0.06	0.083	0	0.36	0.01
Document 2	0.032	0.02	0.24	0.16	0.128	0.04	0.32	0.04	0.02
Document 3	0.06	0.12	0	0.006	0.48	0.174	0	0.08	0.08

The matrix $p(z|d)$ of documents and topics is as follows.

	Art	budget	Education
Document 1	0.9	0	0.1
Document 2	0	0.8	0.2
Document3	0.2	0	0.8

The matrix $p(w|z)$ of topics and terms is as follows.

	movie	music	tax	government	student	teacher	money	art	principal
art	0.3	0.2	0	0.03	0	0.07	0	0.4	0
budget	0.04	0	0.3	0.2	0.01	0	0.4	0.05	0
education	0	0.1	0	0	0.6	0.2	0	0	0.1

With Eq. (11-1), the relationship of these matrixes can be expressed as

$$p(w|d) = \sum_z p(z|d) \times p(w|z)$$

Back to the very beginning, we assume that we have already known the topic distribution of a document and the term distribution of a topic, and then we can generate the words in the document. But it is just on the contrary in reality: we can easily obtain the document set, i.e., the document has already been written. In addition, we have obtained the term distribution of the document, while the unknown part is the topic information of the document, namely, the latent structure of the document. Therefore, according to the document generation process, we have known the results, and now we need to inversely infer the intermediate parameters based on the results, i.e., the topic distribution $p(z|d)$ of the document and the term distribution $p(w|z)$ of the topic. Before describing the calculation method about parameter estimation in details, we will first introduce a formalized presentation and specific algorithm steps of LDA model.

2. Specific Algorithm of LDA Model

Latent Dirichlet Allocation (LDA) model is a hierarchical Bayesian model proposed by David M. Blair together with others in 2003. Here, we formalize LDA model: we assume that the entire document set has T topics, each topic z is expressed as a polynomial distribution θ_z over a dictionary v , and each document d for the T topics has a specific polynomial distributed ϕ_d of a document.

In Figure 11-1, α and β are the parameters for Dirichlet distribution, which are usually fixed values and characterize a symmetric distribution. They usually can be represented by a scalar. θ indicates the topic probability distribution of a document; ϕ indicates the term probability distribution of a topic; and as parameters of a polynomial distribution, θ and ϕ are used for generating topics and words. z represents a topic, w represents a word, M indicates the document number, and N denotes document the length.

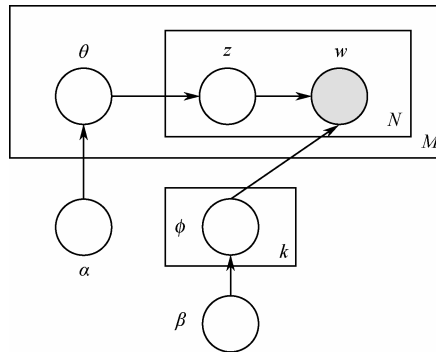


Figure 11-1 Graphical model representation of LDA

Algorithm 11-1 The generation process of LDA model into a document set

- 1: Extract the topic distribution over the term : $\phi \sim \text{Dir}(\beta)$
- 2: From $m = 1$ to M do
- 3: Extract N words: $N \sim \text{Poisson}(\phi)$
- 4: Extract the document distribution over the topic $\theta_m \sim \text{Dir}(\alpha)$
- 5: For $n = 1$ to N do
- 6: Extract a topic $Z_{m,n} \sim \text{Multi}(\theta_m)$
- 7: Extract a word $w_{m,n} \sim \text{Multi}(\theta_{z_{m,n}})$

8: End for 9: End for

During the document generation process in LDA model, firstly the topic term distribution ϕ can be generated (Line 1 in Algorithm 11-1), where ϕ is a parameter in the document level, and it only needs to be sampled once, where the sampling conforms to the Dirichlet distribution with a priori parameter β . For each document, we in the first place determine the document length on the basis of Poisson distribution (Line 3), namely the number of words N . Then, it comes to the step of generating every word in the document. More specifically, a topic can be generated by sampling according to the topic distribution of a document (Line 8), and then a word can be generated by sampling according to the topic term distribution obtained in the former step (Line 9). These steps are repeated until all the words in the document set are generated.

In this process, generally we use the Collapsed Gibbs sampling method to get the values of the hidden variables (z values) and the probability distribution of the parameters in the model (the distribution of θ_m and ϕ). The Collapsed Gibbs sampling process can generally be described as assigning every word in the document set to the corresponding topic. Below we will introduce Gibbs sampling process in detail.

3. Gibbs Sampling in LDA Model

The core procedure of LDA model is the relevant parameter estimation. As parameter estimation is a complex optimization problem, it is very difficult to propose approaches which can obtain precise solutions. Therefore, in general we use approaches which lead to imprecise results, mainly including three ways: ① Gibbs sampling based method; ② calculus of variations based EM solver; ③ expectation advance based approach. Because Gibbs sampling method is simple in reasoning with good results, in practice this algorithm is generally adopted to estimate parameters.

During the process of obtaining the probability distribution of the word, parameters ϕ and θ are not calculated first. Instead, the posterior probability $p(w|z)$ of the word for the topic is considered first and the value of ϕ and θ are obtained indirectly by using Gibbs sampling method. MCMC (Markov Chain Monte Carlo) is a set of approximate iterative methods to extract sample values from complex probability distributions. In this

method, a component of the joint distribution is sampled each time, while the values of the other components remain unchanged. Especially in the case of the joint distribution with higher dimensions, using Gibbs sampling can produce a relatively simple algorithm. As a form of simple implementation of MCMC, Gibbs sampling aims to construct the Markov chain that converges to a target probability distribution, and extracts samples which are considered to be close to the value of the probability distribution from the chain^[27]. As for LDA model in this book, we need to calculate the word distribution over topic, that is, sampling on variable z_i .

Gibbs sampling algorithm can be expatiated as follows^[27]:

(1) As for each i from 1 to N , z_i is initialized to a random integer between 1 to K . This is the initial state of Markov chain.

(2) As for each i from 1 to N , according to the posterior probability calculated by the following formula we assign words to the topic, and get the next state of Markov chain:

$$p(z_{i,j} | z_{-i}, w_i) = \frac{\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + w_\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}}{\sum_{j=1}^T \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + w_\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}}$$

where β can be understood as the frequency of words obtained from topic sampling before seeing any word in the corpus. α can be understood as the frequency of topic sampling before seeing any word in the document. $z_{i,j}$ represents that term w_i is assigned to the topic j , and z_{-i} represents the distribution of all $z_k (k = i)$. $n_{-i,j}^{(w_i)}$ is the number of words assigned to topic j which is the same as w_i ; $n_{-i,j}^{(\cdot)}$ is the number of all words assigned to topic j ; $n_{-i,j}^{(d_i)}$ is the number of words assigned to topic j in document d_i ; and $n_{-i,\cdot}^{(d_i)}$ is the number of all words that are assigned with topics in d_i . All the numbers of words exclude the allocation of $z_{i,j}$ this time.

(3) After a sufficient number of step (2) of iterations, we can believe that Markov chain approaches the target distribution, then take the current value of z_i (i from 1 to N) as the sample recorded. In order to ensure the autocorrelation smaller, we need to record other samples after each certain times of iteration. Abandon word mark and set w represent words. For every single sample, estimate values of ϕ and θ according to the following formula:

$$\tilde{\phi}_w^{z=i} = \frac{n_j^{(w)} + \beta}{n_{\cdot}^{(\cdot)} + w_{\beta}}, \tilde{\theta}_{z=i}^d = \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{(d)} + K\alpha}$$

where $n_j^{(w)}$ represents the frequency of words w assigned to topic j ; $n_j^{(\cdot)}$ represents the number of all words assigned to topic j ; $n_j^{(d)}$ indicates the number of words assigned to topic j in document d ; and $n_{\cdot}^{(d)}$ represents all the number of words which have been assigned with topics in document d .

Gibbs sampling algorithm starts from an initial value, and after iterated for enough times it can be considered as the probability distribution of the sample closed to that of the target.

Here we use a simple example to illustrate the iterative process of Gibbs sampling. The selected five documents are from the Facebook about MH370 flight lost topic (for details, see Appendix), and their time is 20140309, 20140323, 20140325, 20140327, and 20140405, respectively. When the number of topics is set to be 5, we get a series of Gibbs sampling results as follows.

	Iteration 1600 times	Iteration 1700 times	Iteration 1800 times	Iteration 1900 times	Iteration 2000 times
Topic 1	8.73	9.78	9.28	9.31	9.12
Topic 2	9.98	9.86	9.6	9.92	10.03
Topic 3	10.04	10.8	10.99	10.99	11.51
Topic 4	10.52	10.4	10.71	10.92	10.64
Topic 5	9.11	9.19	9.23	9.45	9.86

4. Application Examples of LDA Model Based Topic Discovery

Although the theory of LDA model looks relatively complicated, since LDA theory is mature and widely used, there are many LDA model implementation codes available for free downloading on the Web. In the application process of LDA, we can take full advantage of existing programs or tools and concern only with the results of the input and output without focusing too much on the computational details. For example, in the application example of this section, the code of LDA model we used is Mallet Kit of UMass (MACHine Learning for Language Toolkit)^①, using the default parameter settings.

In the examples in this section, by using the five documents adopted in above example of the Gibbs iterative process and implementing the algorithm, we get a series of results

① <http://mallet.cs.umass.edu/>.

such as the relationship of topic and terms in a document, which are as follows.

Next we will take a document on Facebook from 20140323 as an example to illustrate the results obtained by LDA model. In this document, we use different underscores to represent the different topics (i.e., Topic1 Topic2 Topic3 Topic4 Topic5). After underlining the following documents, we can note which topic each word belongs to (words not-underlined are prepositions or other words helpless in understanding the document, which should be ignored during processing).

MH370: Indian aircrafts join search

KUALA LUMPUR: Two long range maritime reconnaissance aircrafts from India have departed the Subang Airport on Sunday to join in the search for the missing Malaysia Airlines (MAS) MH370 aircraft in the southern Indian Ocean.

This was despite reported bad weather caused by tropical cyclone Gilia which had forced a number of other aircrafts scheduled to go on the mission to be cancelled.

The High Commission of India to Malaysia on Sunday said that both aircrafts were likely to encounter the cyclonic conditions enroute, but Captains of both aircrafts had instead “decided to skirt bad weather areas” to reach the search sectors.

The two aircrafts were the P8-I of the Indian Navy and the C-130J of the Indian Air Force. Both were expected to undertake a 10-hour mission in the area.

“Both aircrafts have long endurance capabilities coupled with state of the art electro optronic and infra red search and reconnaissance equipment on board,” said the high commission in a statement here.

The P8-I aircraft has the added advantage of on-board radars and specially designed search and rescue kits.

Previously, the Prime Minister of India had committed to assisting Malaysia and “render all possible assistance to Malaysia in locating the missing Malaysian Airlines flight”.

Both aircrafts arrived in Malaysia on March 21. After extensive briefings in Malaysia on Sunday, both Indian aircrafts took off to be part of the Australian-led search on Sunday morning.

India has been participating in the search and rescue operation beginning March 11 in the Andaman Sea and Bay of Bengal.

By the proportion of various underlined words in the document, we can see the distribution of various topics in this document, that is, a large proportion of some underline

words indicates that its corresponding topic takes a large proportion in the document.

Besides, we can obtain the following topic term distribution graph (As Mallet just gives the weight of each term, Figure 11-2 is not the real distribution graph but only depicts the renormalization results of each word corresponding to the weight in each topic, but it still indirectly reflects the distribution of the topic terms), where the horizontal axis represents each term (because there are a lot of terms, we have not marked them in the figure), and the vertical axis represents the weight of each term. From Figure 11-2 we can intuitively observe the difference between different topics, which is reflected by different term distributions.

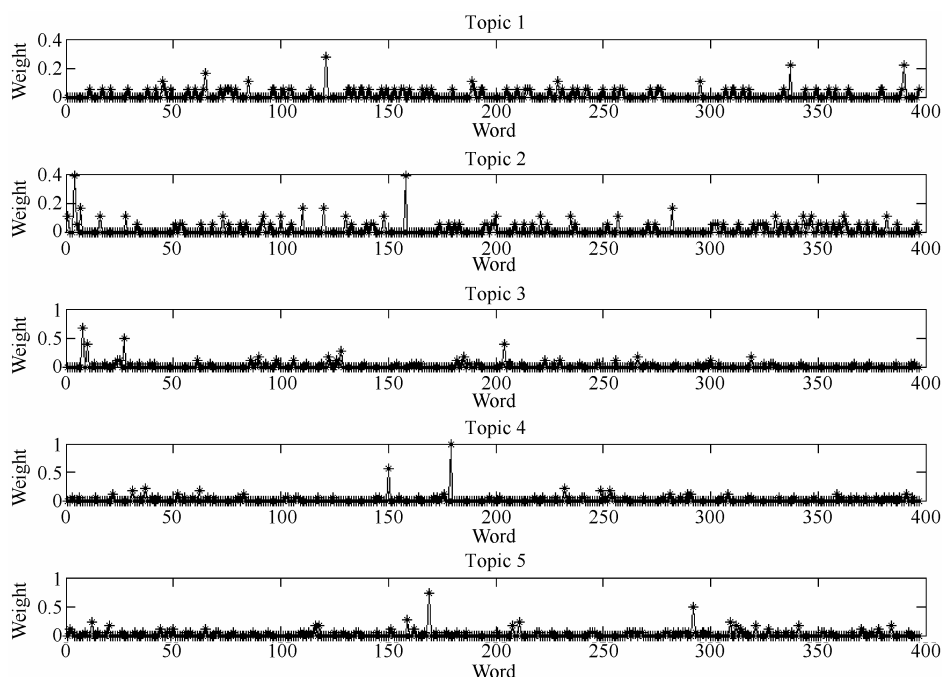


Figure 11-2 Topic term distribution graph

The above examples with underlines provide an intuitive feel for topic distribution, and more specific values can be found in Figure 11-3, where it reflects the topic distribution of each document in the form of a line chart. Documents 1, 2, 3, 4, and 5 refer to five documents aforementioned. As the five documents are in chronological order, Figure 11-3 can simply be seen on the evolution situations of the topic of MH370 aircraft lost event (more specific analysis of topic evolution will be described in detail in Section 11.3).

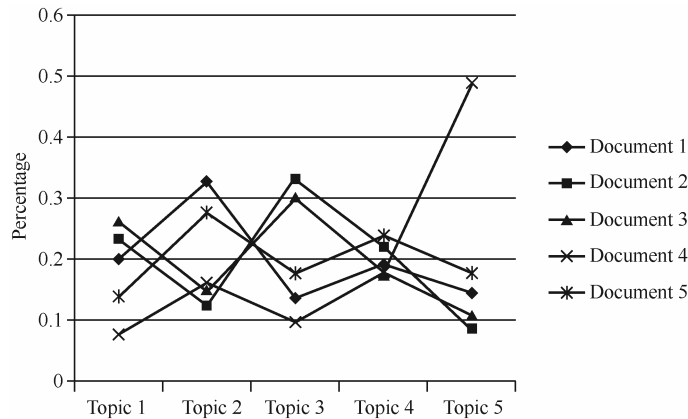


Figure 11-3 The distribution of different topics in each document

5. Research on Topic Model in Topic Discovery of Social Networks

Topic model is mainly used to discover the latent semantic structure in documents, i.e., the abstract topics in documents. Obviously, we will tend to apply the topic model to the topic discovery of contents in social networks. But due to certain features of social network data, relevant researches show that the direct application of the topic model to social networks (such as short text data) does not produce the expected results^[17]. Therefore, we need to study how to use the topic model in the social networking environments.

Social network data includes the data from blog, microblog (e.g, Twitter, Weibo, etc.), instant message (IM, such as QQ, etc.), etc. Since twitter was launched in 2006, it has covered more and more users, and thus, it has gradually become the most popular data source for social network researchers. With the data from twitter, the researchers have done a lot of work. Particularly, they try to use topic model on tweets, and intent to find more valuable latent topic information.

Aiming at the problem that due to the content length limit of tweets the direct application of traditional topic model is ineffective, some scholars study how to train the standard topic model in the short text environments. They combine tweets with the authors' information, and present three topic modeling models^[13].

(1) MSG Model: MSG refers to Messages Generated by the Same User.

- ① Train LDA model in training corpus;
- ② In the training corpus, the information generated by the same user is aggregated

into a Training User Profile for the user;

③ In the testing corpus, the information generated by the same user is aggregated into a Testing User Profile for the user;

④ Take Training User Profile, Testing User Profile and testing corpus as “new documents”, and use the training model to infer the distribution of their topics.

(2) USER Model:

① In the training corpus, the information generated by the same author is aggregated into the User Profiles, in which we train LDA model;

② In the testing corpus, the information generated by the same author is aggregated into the Testing User Profile for the user;

③ Take training corpus, Testing User Profile and testing corpus as “new documents”, and use the training topic model to infer the distribution of their respective topics.

Obviously, MSG model and USER model are not suitable for topic modeling of a single Tweet but can be applied to detect the topic distribution of Tweets posted by a specific author. Both models use the user based aggregation policy, but the order and the manner of training models are different.

(3) TERM Model: As the name itself suggests, it is the aggregation of all information that contains a certain term.

① For each term in the training set, all information containing the term is aggregated into a Training Term Profile;

② Train LDA model on the Training Term Profile;

③ Establish User Profiles (namely aggregate the information issued by the same user) on the training set and testing set respectively;

④ Take training corpus, Training User Profiles, Testing User Profiles and testing corpus as “new documents”, and use training models to infer the distribution of their respective topics.

The principle of TERM is based on the fact that twitter users often use customized topic labels [the words surrounded with # called tweets label (Hashtag)] which represents a specific topics or events. With the established Term Profiles, using TERM model can directly obtain topics related to these topic labels.

MSG model trains LDA model by a single Tweet. As the length of the content itself is limited, there is not enough information for the model to learn topic pattern. Specifically, compared with the long text, the words of short text have a lower distinction for text. The TERM mode and USER mode, however, use the aggregation strategy to train the model,

and the results they obtained should be better.

For the difference between the content of twitter and that of traditional media, some scholars have proposed an improved LDA model — twitter-LDA^[32]. Clearly, because tweets are brief, the standard LDA model does not work well on twitter. In order to overcome this problem, the above mentioned MSG, USER and TERM three modes use aggregation strategy to merge the same user's tweets to a longer article, which, by and large, is the application of the Author-Topic Model (ATM)^[21]. But this can only discover the same author's topic, which is useless for a single tweet. Hence, we have to improve the standard LDA model so that it can form a model which is useful for a single tweet.

In twitter-LDA model, we assume that a tweet has K topics, and each topic can be represented by a distribution of terms. φ^t represents the term distribution of the topic t ; φ^B represents the term distribution of background words; θ^u represents the topic distribution of the user u ; and π represents a Bernoulli distribution (used to control the choice between background words and topic words). When writing a tweet, a user firstly selects a topic based on its topic distribution, and then chooses a group of words according to the selected topic or background model. twitter-LDA model can be shown in Figure 11-4, with the generation process as follows:

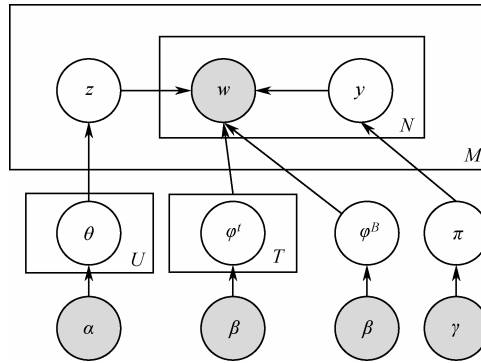


Figure 11-4 Twitter-LDA model

- (a) Sampling $\varphi^B \sim \text{Dir}(\beta)$, $\pi \sim \text{Dir}(\gamma)$
- (b) For each topic t :
 - a) Sampling $\varphi^t \sim \text{Dir}(\beta)$
- (c) For each user u :
 - b) Sampling $\theta^u \sim \text{Dir}(\alpha)$
 - c) For each Tweet &

- i . Sampling $z_{u,\&} \sim \text{Multi}(\theta^u)$
- ii . For each word n in $\&$
 - A. Sampling $y_{u,s,n} \sim \text{Multi}(\pi)$
 - B. If $y_{u,s,n} = 0$, sampling $w_{u,s,n} \sim \text{Multi}(\varphi^B)$
 - If $y_{u,s,n} = 1$, sampling $w_{u,s,n} \sim \text{Multi}(z^{u,s})$

Twitter-LDA is actually the extension of ATM, more specifically, the background knowledge is introduced into the ATM model. Twitter-LDA models both users' background and the background of tweets. Hence, to some extent, it overcomes the limitations of ATM which only models users' background.

Another variation of topic model used on twitter is Labeled LDA ^[20], which expands the LDA model and combines the supervision information. Labeled LDA assumes that there is a set of tags \mathcal{A} , where each tag is represented by a polynomial distribution $\beta_k (k \in 1 \cdots |\mathcal{A}|)$ of a term. Each document d only uses a subset of tags \mathcal{A} , labeled as $\mathcal{A}_d \subset \mathcal{A}$, and document d represents a polynomial distribution of \mathcal{A}_d . Labeled graph model is shown in Figure 11-5, with the generation process described as follows:

- (a) Sampling the topic distribution $\beta_k \sim \text{Dir}(\eta)$
- (b) For each tweet d :
 - a) Establish a label set \mathcal{A}_d in describing the Tweet d from a hyper parameter ϕ
 - b) Select a polynomial distribution θ_d from the label set \mathcal{A}_d according to the symmetrical prior parameter α
- c) For each lexeme i in tweet d
 - i . Sampling labels $z_{d,i} \sim \text{Dir}(\theta_d)$
 - ii . Sampling terms $w_{d,i} \sim \text{Dir}(\beta_z)$

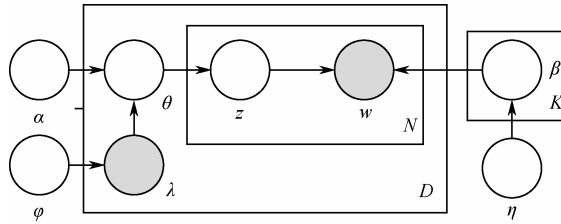


Figure 11-5 The graphical model as the representation of Labeled LDA

To adopt Labeled LDA to analyze the features of twitter content ^[19], we can relay on the basis of the features of twitter contents and regard each hashtag as a label in tweets,

which to some degree uses the topic information that users tagged. But the drawback of this approach is obvious that the model can't be directly applied to all tweets without any hashtag.

In addition to the twitter data, some scholars have studied the chat data from the Internet Relay Chat (IRC) ^[29]. Just like twitter data, IRC data also includes misspellings, grammatical errors, abbreviations and other noise information. Along with such features as dynamic change, concise expression, intersected topic, etc., IRC data is not suitable for us to analyze with the existing text mining methods. In order to remedy these problems, researchers use the latent features of IRC data — social relationships among chatters, to filter out the irrelevant parts in discussion. This is similar to the use of PageRank algorithm that distinguishes highly interacted web pages from irrelevant web pages. However, it should be noticed that the score here is not for ranking but to improve the topic model via such correlation score. This method can be summarized as using some features of IRC data to construct user relationship graph, and then according to the features of the graph (such as indegree, outdegree, etc.) provide a score for the user, which is used to determine which is noise information and which is information related to the topic. Additionally, in accordance with relevant scores, noise information is decreased and useful information is enlarged so that the application effect of the topic model can be improved.

Topic model has got a significant amount of attentions from scholars in various fields as it not only has a solid statistics theoretical foundation, but also is simple in model and easy to be extended to accommodate different applications. Many scholars try to apply topic model on social networks. Besides some of the models mentioned above, they also propose a variety of other topic models, which provides an effective means for us to analyze social network data.

11.2.2 Vector Space Model Based Topic Discovery

VSM (Vector Space Model) ^[23] was originally proposed by Gerard M. Salton et al. in the 1970s, and it was successfully used in information retrieval field. The basic idea of the VSM model is that the document or query is represented as the form of eigenvector, so that the processing of the text content can be transformed into vector operations in vector space, and the similarity of vectors reflects the similarity between documents. More specifically, firstly the features of these texts are usually represented as terms (the specific differences between term and word can be accessed in information retrieval section of this book). In

addition, TFIDF value represents the weight of the feature. Finally, the similarity between vectors can be measured with a cosine function. This method is relatively simple, so it will not be introduced in detail here.

In the initial TDT project, the topic discovery was understood as the classification of a large number of different news stories; therefore, the methods focus on the Vector Space Model (VSM) based clustering method. As for different application scenarios, researchers propose different solutions. As for Retrospective Event Detection (RED), scholars have proposed a Group Average Clustering (GAC) algorithm; As to online event detection, a Single Pass Incremental Clustering (SPIC) algorithm is proposed. These methods well adapt to the needs of different scenarios and meet task requirements.

The adoption of the vector space model for topic discovery is based on the assumption that the documents of similar topics are similar in content. Thus, after transforming the document as eigenvector by VSM, the key step is to measure the similarity between features. However, the clustering algorithm used in traditional TDT mainly aims at the topic discovery in the data stream of news report, and migrating this method directly to the current social network environment has limitations. This is due to the fact that in Twitter, Weibo and other similar sites the data generated by users are relatively short with a frequent use of informal languages and there are specialized network terminologies etc., which makes the VSM model that simply feature terms produce sparse data and other issues. Consequently, we need to improve VSM method to make it applicable to the data generated by social networks. When constructing eigenvector with the VSM model, it is necessary to pay much more attention to feature selection, the calculation of feature weight and similarity measurement between features. In simply words, the chosen features and the calculation of feature weight should be sufficient to represent the content of current text and reflect the difference between the texts. The similarity measurement should accurately reflect the difference between features.

As for the problems of feature selection and the weight calculation of text contents in social networks, researchers have done a lot of work. Here we briefly introduce these efforts.

Paige H. Adams and Craig H. Martell compare the effect of several feature selection and weight calculation methods for the similarity measurement of Chat (i.e., instant messaging) data in the reference [1]. Authors analyze Chat data and obtain the following characteristics:

- (1) Topics tend to aggregate by time. New topics may come from the previous topics;

before dying or replaced by new topics, the current topic will remain for a while.

(2) Interleaving occurs between different topics. Messages containing different topics may be mixed together.

(3) The participants of a specific topic may change. But generally speaking, the core participants of a topic dialogue tend to remain unchanged.

Based on the above data characteristics, the authors point out that it is necessary to improve the traditional weight calculation method of using TFIDF feature:

As for feature (1), during the similarity measurement, the Time Distance Penalization coefficients are introduced in order to increase the similarity of data with similar time, and reduce the similarity of the data with larger time difference.

As for feature (2), with regard to term–featured aspect, the authors use Hypemym (i.e., topic words of conceptually broader denotations) to solve the problems that the different representations of the same topic result in high semantic similarity but low coincidence degree of terms.

As for feature (3), the author assigns each eigenvector with user’s nickname information corresponding to the original document and gives the data published by the same user more weight, which implies that the information published by the same author has a higher probability under the same topic.

Hila Becker et al. analyze the rich context (i.e., background or contextual content) information in social networks in the reference [3], including text messages and non-text information, such as user labeled information (such as title, tags, etc.), automatically generated information (such as creation time of content, etc.) and so on. With the aid of the above background information, the authors compare the various measurement techniques for the similarity of documents from social media. In this paper, the authors point out that the single feature of social media content remains with big noise and it can’t be used to effectively complete classification and clustering task, but the joint use of various features (including the document content and context of content, etc.) can provide us valuable information related to the topic or event.

In fact, although context features between different social medias are not the same, many social media still share some of the same features, such as author name (i.e., the user of document creation), title (i.e., the name of the document), description (i.e., a brief summary of the document content), label (i.e., a set of keywords describing the content of the document), date and time (i.e., the time of content published) and location information. These features can provide help to measure the similarity between documents. For different

features, researchers provide different approaches:

(1) As for textual features, the eigenvector identifying with TFIDF as weights can be used to calculate the similarity by a cosine function.

(2) As for date and time features, the similarity is calculated as $1 - \frac{|t_1 - t_2|}{y}$, where t_1 and t_2 respectively represent the time of publishing of document 1 and document 2, and y shows the number of minutes per year. In addition, if the time interval of two documents is more than a year, then the similarity of the two documents is regarded as zero.

(3) As for geographical features, the similarity is calculated as $1 - H(L_1 - L_2)$, where L_1 and L_2 respectively represent the latitude and longitude of document 1 and document 2, and $H(\bullet)$ function is to use a Haversine distance.

In the task of topic discovery based on clustering algorithm, the key point is to determine appropriate similarity measurement approaches to reflect the similarity between documents. In regard to similarity measurement methods, Hila Becker et al. put forward two ways in the reference [3], including the aggregation based similarity measurement method and the classification based similarity measurement method, among which the former adopts a variety of clustering methods, and the final result is determined by weighted voting based on different weights of clustering methods in the clusterers or weighted combination in the calculation of the similarity. The idea of classification based similarity measurement takes similarity score as classification feature for predicting whether two documents belong to the same topic or event.

After considering feature selection and similarity measurement, we also need to explore the selection of algorithms including clustering. Most of the data in social networks can be seen as the continuous incoming data stream, which characterizes in large-scale and real-time. Hence, in the choice of clustering algorithms, we must choose a clustering algorithm which can be expanded and requires no prior knowledge of the number of clusters ^[4]. Thus, researchers have proposed an incremental clustering algorithm, which takes into account each piece of information in turn and decides the appropriate category based on the similarity between the information and the current clusters.

Here we briefly introduce a single-pass incremental clustering algorithm that only scans the data once and dynamically adjusts the parameters, perfect for online environment where there are continuous data and dynamic increase of the number of categories of clustering. The algorithm steps are as follows:

Algorithm 11-2 Single-pass incremental clustering algorithm

```

1: Set two thresholds: clustering threshold  $t_c$ ; creating a new type of threshold  $t_n$ 
2: The input of a document  $x$ 
3: For 1 to  $T$  do:
4:   Calculate the similarity  $\text{sim}(x, t)$  between  $x$  and the current topic  $t$ 
5:   If  $\text{sim}(x, t) > t_c$  do:
6:     Refer  $x$  as the current topic  $t$ , and add into the document cluster of  $t$ 
7:     Recalculate the center of the topic  $t$  (a document cluster), which
       represents topic  $t$ 
8:   end cycle
9:   end If
10:  If  $t_n < \text{sim}(x, t) \leq t_c$  do:
11:    Without any treatment, continue to cycle
12:  end If
13:  If  $\text{sim}(x, t) \leq t_n$  do:
14:    Create a new topic  $t'$ , whose center point is represented by  $x$ ,
       and add it into the topic set  $T$ 
15:  end cycle
16:  end If
17: end For

```

Single-pass incremental clustering approach measures the similarity between each incoming data and each cluster (line 4). More specifically, if the incoming data is sufficiently similar to a cluster, it can be classified into this cluster (line 6); if it is not similar to any of the current clusters, we have to create a new category and put the incoming data into it (line 13). The center representing the cluster can be represented by a vector formed by the mean value of each dimension of the cluster.

In summary, the VSM based topic discovery method is with the traditional clustering ideas, but as for some of the new features in social networks, researchers have put forward improved scheme in feature selection, weight calculation and similarity measurement and other aspects, which solves the encountered problems and improves the results of topic found to some extent.

11.2.3 Term Relationship Graph Based Topic Discovery

Term co-occurrence analysis is one of the successful applications of natural language processing techniques in information retrieval. It is established in the idea that the frequency of co-occurrence between terms to some extent reflects the semantic relation between terms. Scholars at the very beginning make use of term co-occurrence to calculate the similarity of the documents, and then they apply this method to complete topic word extraction, topic sentence extraction, summary generation and other tasks.

The method of term co-occurrence is mostly based on such an assumption: in corpus sets, if two terms frequently appear in the same document, the combination of these two terms is considered to be stable and semantically interlinked, where the frequency of co-occurrence reflects the closeness of semantics between terms. It can be further expected by such assumption that a topic or an event can be represented by a series of descriptive and parallel keywords, while according to intuitive idea, the same set of keywords tend to be used to describe the same topic or events among different documents, based on which the links between keywords will be closer, and the frequency of co-occurrence will be higher. In short, the closer the relationship between the term co-occurrence under the same topic is, the more the chance of co-occurrence is; In contrast, the weaker the co-occurrence relationship is between terms under different topics, the less the chance of co-occurrence is. Thus, by the relationship of co-occurrence between terms in a document, we can reversely find out which words are related topics (used to describe the same topic) so as to achieve the purpose of topic detection.

Based on the above ideas, the term co-occurrence based approach is turned out to be useful on topic discovery. The term co-occurrence relationship can be shown in graph form. The terms of close relationship can be intuitively found from the graph and further forms a set of terms related to the topic, which has become a simple and effective method in topic discovery.

1. Basic Idea

The basic idea of the term relationship graph based topic discovery algorithm is firstly to construct a term co-occurrence graph based on the co-occurrence relationship between terms, then via the community detection algorithms commonly adopted in social network analysis to find the community formed by terms (i.e., a set of terms related to a certain topic), and finally to determine the topic of each document according to the similarity

between the found term sets and the original document.

2. Method Description

The term co-occurrence graph based topic discovery can be largely divided into the following three steps^[25,26]:

Step 1, construct a term co-occurrence graph according to the relationship of term co-occurrence;

Step 2, conduct a community detection algorithm in the term co-occurrence graph, which leads to a community (i.e., term set) with the description for a specific topic;

Step 3, specify a set of topic terms for each document in the original document set.

Next, we will conduct a detailed description for each of the above steps.

1) Step 1: Construct the term co-occurrence graph

In the term co-occurrence based topic discovery method, terms are generally taken as nodes in the graph. In the term co-occurrence based topic discovery method, terms are generally taken as nodes in the graph. However, due to the fact that the importance of different terms differs, it is unnecessary to construct a term co-occurrence graph with all words appeared in the document, instead, the keywords in the document should be regarded as the nodes, which is because the keywords of a document can better distinguish topics for a document. The keywords can simply be identified by using document frequency (DF), and the words with the document frequency below a certain threshold $node_min_df$ should be deleted^[18,25,26].

After selecting terms as nodes, we start to construct edges between nodes and the edges in the term co-occurrence graph should reflect the co-occurrence relationship between terms, namely, if two terms co-occur at least in a document, then an edge between the two terms (nodes) can be established.

After constructing the term co-occurrence graph, we need to pre-process the generated graph to remove some edges to reduce the size of graph and remove noise information. Detailed process can refer to the following rules:

Now we need to calculate the total number of term co-occurrence (the number of documents with two term co-occurrence, here is equivalent to the document frequency of edges in the term co-occurrence graph), and referred as $DF_{e_{i,j}}$, representing the co-occurrence frequency of term w_i and w_j appeared jointly in the document set, and also representing the document frequency of edge $e_{i,j}$. We firstly remove edges whose co-occurrence frequency is below the threshold $edge_min_df$.

As for edge screening, we can also calculate the conditional probability of the emergence

between terms, and then remove the edge between corresponding nodes when the corresponding bidirectional conditional probability is below the threshold `edge_min_prob`. Conditional probability is calculated as follows:

$$p(w_i | w_j) = \frac{DF_{i \cap j}}{DF_j} = \frac{DF_{e_{i,j}}}{DF_j}$$

where DF_j represents the document frequency of term w_j ; $p(w_i | w_j)$ represents the conditional probability that term w_i appears at the same time the term w_j appears; and $p(w_j | w_i)$ is defined similarly.

2) Step 2: Perform community discovery algorithm in the term co-occurrence graph

As mentioned earlier, we can make such an assumption that if there is topic relation with the same meaning between two words, the two words will appear together with a higher probability. Based on such an assumption, we construct the co-occurrence graph between terms and we also have formed a co-occurrence based term network from the previous step. In such a network, the terms used to describe the same topic connect closely, while the terms applied to describe different topics are the opposite. From this we can adopt the ideas of community discovery in social networks and divide the term co-occurrence network into “communities” — represents a set of terms corresponding to specific topics — to describe different topics.

Community discovery may draws support from betweenness centrality to discover the connected edges between two communities. This method is based on the intuitive understanding that when calculating the shortest path of nodes between two different communities, it is inevitable to pass through the edge connecting two communities, and then for this type of edge, its betweenness centrality value is relatively high. Therefore, by calculating betweenness centrality, the edges extending transversely across communities can be found. Via removing the edge with higher betweenness centrality values, equivalent to cutting off the path between communities associated with the edge, the task of community discovery, i.e., the topic discovery, is achieved.

If the correlation between topics is weak, that is, no basic connection between different topics, there is no edge connected between different topic term sets and then we can even use non-connectivity sub-graphs to discover topics directly.

Of course, there are other algorithms for topic cluster discovery, for example, for every term, it is firstly placed in a community; and for a neighboring node of the current community, if its connection with the current community exceeds a certain threshold, it can

be added to the current cluster. Repeat the process until no additional nodes added to the current community, and then the community formed by the current terms is a term set describing a specific topic.

Definitely, the above topic clustering discovery algorithm is only an example for illustration. Because here the problem of topic discovery is converted to a community discovery task, theoretically the community discovery algorithm that can effectively detect a term set related to a topic can all be applied to the topic discovery.

3) Step 3: Set topic marker to the original document set

After determining the topic cluster, we need to make a topic judgment about the original document sets, namely to determine the relationship between each document and the topic cluster. Intuitively, for each topic cluster, the more parts of the topic cluster appeared in a document, the more relevant the topic associated with the topic cluster in the document. In this step, we consider each term in the topic cluster as a feature of the topic, and then they can form an eigenvector of the topic. Our task is to calculate the similarity μ between the cluster and each document, and the similarity calculation can be simply represented by intersection with the calculation formula as follows:

$$\mu = \frac{\sum_{w \in d \cap S} f(w)}{\sum_{w \in S} f(w)}$$

Here we define the topic cluster as S and the document as d , and $d \cap S$ represents the intersection of the document and terms in topic clusters. $f(w)$ can be a simple Boolean or other functions, and its value reflects the similarity between the topic cluster and the document.

Certainly, we can use a cosine function to measure the similarity between a topic and a document, and hence the probability distribution of each topic in document d can be calculated as:

$$p(z | d) = \frac{\cos(d, S_z)}{\sum_{z' \in Z} \cos(d, S_{z'})}$$

In general, the different features of the topic have different weights on this topic, so we can also on the basis of this feature improve the similarity measurement method for a topic and a document. For example, we can use TF*IDF value to evaluate the weight of different features (i.e., terms) under the current topic so as to obtain a more accurate topic

distribution in the document.

Overall, the term co-occurrence graph based topic discovery adopts a more matured term co-occurrence theory, which is intuitive and simple, providing new ideas and methods for topic discovery.

In summary, topic discovery technology originates from TDT project of DARPA ^[2], originally using VSM model based clustering methods. Later, with the development of topic models, scholars gradually begin to conduct text analysis with topic models. Meanwhile, the other methods such as term co-occurrence analysis also stand as pretty good ideas. Definitely, not all topic discovery methods are mentioned in this chapter, for example, we can also use some of the new technologies from natural language processing. For instance, the recent emerging of deep learning also provides a new technical means for topic discovery. In another example, focusing on extensive user participation and rich interactive features in Weibo social media, a group intelligence based new topic/event detection method for Weibo streams is proposed in the reference [9]. The main idea of this approach is firstly to decompose the traditional single mode topic/event detection model according to Weibo participants, then create a language feature and sequential characteristic based topic/event discriminant model for each participant, namely the Weibo user personal model, and finally determine the new topics/events by voting with group intelligence. As we have mentioned earlier, for the various types of data, we cannot simply transplant these technologies and methods. Instead, we should analyze the features of these data and modify the traditional methods properly, only in this way can the effect of the adopted method and the accuracy of topic discovery be improved.

With the explosive growth of social networks, new social media represented by Weibo will surely get more widespread concern. Analyzing the content of social networks, knowing better their users' interests and providing valuable information for business decisions and public opinion monitoring can all find root on the topic discovery technologies in social networks. The rapid development of social networks provides plentiful research materials for topic discovery, which in turn requires topic discovery technologies advance with times, and adapt to the emerging new media and new data. In short, the prospect of technology application of topic discovery is bound to be more and more broad with the pace of the Internet age.

11.3 Models and Algorithms of Topic Evolution

Information in social networks is continuously updated due to the dynamic characteristics of data in social networks. In this case, how to track the development trend and future development of the topic interested to users becomes a key problem of user's concern and a key problem to be solved. With advance of time, the content of a topic in social networks may change accordingly, and a topic strength may also undergo a changing process from a low tide to a high tide or backwards. Consequently, how to effectively keep detecting topics in social networks and obtain topic evolution in chronological order so as to help users of social network track topics has a very significant realistic demand and a practical value. Especially in public opinion monitoring in social networks, timely and effectively tracking the evolution situation of sensitive topics and making an appropriate forecast are core requirements for sensitive topic monitoring, which has broad application prospects and important application values.

In the early TDT research^[2], the main consideration of topic tracking is the dynamics, development and difference of a topic as time flies and the main technical approach is to use statistical knowledge to filter text information, followed by the adoption of classification strategy to track relevant topics. However, these earlier studies have not effectively considered time characteristics of terms and analyzed the distribution of topics across the timeline.

With the topic model proposed, how to effectively use the time characteristics of terms in the topic model and study the characteristics of topic evolution becomes a hotspot issue in study of social networking text. Unlike the earlier TDT study, each text in the topic model is a mixed distribution of topics and each topic is a mixed distribution of a set of terms. Because the topic model can capture the evolution of a topic and the introduction of topics may have a good effect on text prediction, the topic model has been widely used in the field of topic evolution.

This section focuses mainly on the most common and the most widely-used topic evolution methods in social networks. Firstly, we will introduce the most widely-used simple topic evolution method. Secondly, we will introduce the LDA model and other topic models with high precision which can meet various bonding time of a variety of application requirements. Finally, topic evaluation methods with better application effect in some special application environments will be introduced.

11.3.1 Simple Topic Evolution

In the study of the topic evolution in social networks, the most commonly-used method is the simple topic evolution method: using a topic discovery method in each time slice, and then analyzing and comparing the similarity of keywords obtained by the topic discovery algorithm of the adjacent time slices in order to analyze the situation of topic evolution^[27].

A typical article^[25] about topic evolution points out that due to the fact that the social network continues to generate large amounts of data, it becomes impossible to perform topic discovery algorithm on the entire data set after each generation of new data. So they set a sliding time window with a fixed size in social networks, and use their proposed topic discovery algorithm to calculate the similarities and differences between the topic of this window and the topic of the previous window for analyzing the evolution situation of the topic.

There are many proposed topic evolution methods of the topic expression based on multiple words and supervised/semi-supervised algorithms which can solve the problems of topic evolution in traditional media. But in the study of topic evolution in social networks, these traditional methods have the following problems:

- (1) There is more noise in social network text than that in traditional text, such as colloquial text, slang, advertising, etc.
- (2) The text in social networks is shorter than that in traditional media, which make the accuracy of text mining algorithms greatly reduced.
- (3) The topic in social networks is word of mouth, which makes the topic evolve very fast.

Taking into account the above-mentioned characteristics of the topic evolution in social networks, Tang et al. puts forward a semantic graph model in the reference [28] to study the topic evolution. In this method, as for the above mentioned problems of semantic sparse, much noise and short text, they introduce the existing knowledge base (such as Wikipedia) to solve these problems. Specifically, for any published blog, they take the name of the entity and concepts on Wikipedia as nodes, using the relationship between nodes and semantics measured by graph edit distance (GED) as the weight of the link to construct the semantic graph. Irrelevant concepts and noise will be filtered out by a graph clustering algorithm in the semantic graph. The model can be updated in line with the

dynamic update of blogs to track the topic evolution. Moreover, this model also has the following characteristics:

(1) This semantic information based model can solve the problem of synonymous. As traditional word classification methods cannot distinguish the semantic similarity of words, we cannot find synonyms that express the same topic. In contrast, in the Wikipedia, synonyms are clustered together by links and all synonyms can be effectively mapped to the same concept.

(2) Graph clustering algorithm will effectively filter the noise and multi-topic texts. By analyzing the semantic distance between keywords, graph clustering algorithm will discover the main topic of a blog.

(3) There is no requirement for statistics on the training sets used in statistical methods because the graph edit distance is applied in this method.

(4) Since this method uses the similarity between semantics, this method is particularly suitable for topic evolution.

The evolution analysis of public concerns is a variation of topic evolution analysis. In the case of Weibo, there is not only Weibo describing event information, but also Weibo reflecting public concerns. The event information in Weibo reflects the occurrence and development of events, while public concern in Weibo embodies the public's interest and expectations, opinions and attitudes, and emotions and tendencies for events. For example, in a commercial event, public concern reflects their demand for products and their evaluation for after-sale service and brand reputation. Therefore, understanding and analyzing the public concern become an important part on mastering Weibo events.

In the existing research on analyzing the public concerns in events, firstly we need to predict and specify the particular side of events required to observe, so this method is not suitable for unknown general event. Deng et al. take the long Weibo and the forward and comment Weibo as research object in the reference [10], and propose an unsupervised evolution analysis of public concern to reconstruct topic space with long Weibo, then map forward and comment Weibo to the space to perform correlation analysis with long Weibo and transform the evolution analysis of public concern into the position tracking of the forward and comment Weibo in the topic space. Therefore, under the situation without predicting public concerns in events, we can portray the evolution process of public concerns with the development of events. Interested readers can refer to the reference [10] for more details.

11.3.2 Topic Model Based Topic Evolution

Topic model as the most popular and commonly-used approach in topic discovery also has very important applications in topic evolution in social networks.

1. LDA Based Topic Evolution

With respect to the topic research in online social networks, reference ^[7] proposes online LDA topic evolution with the basic idea as follows: firstly, a sliding window technique is adopted to divide the text stream of social networks into time slices; and then the LDA model is used to model the document within each time slice, where each document is represented as mixed topics with LDA and each topic is a polynomial distribution of pre-set words, which leads to the probability distribution of Topic - Word and Document - Topic. Moreover, the authors use the relationship between the posterior probability and the priori probability to maintain the continuousness between topics, namely, adding a weight W to the Topic—Word probability distribution of the previous time slice as the prior probability of the current time slice and establishing an online LDA calculation model. Under the guidance of KL similarity metric distance measurement, according to the changes of probability distribution of Topic - Word and Document - Topic over time, there are two fundamental topic evolution models: the topic evolution and the topic strength evolution.

Due to the fact that there is no requirement for data of the time slice when dealing with the current time slice, online LDA model thus saves the memory to handle large-scale corpus, which is suitable to online social network environments.

LDA model is also applied in another social networks – the topic evolution in research citation network. In research citation network, the reference naturally reflects the relationship between topics. Bolelli et al.^[6] in 2009 first introduce the LDA in the research of research citation network, but they simply use references to identify and determine the weights of a few words that describe topics most properly. In reference [11], He et al. propose a systematic analysis approach for topic evolution in research citation network. Firstly, the authors extend the LDA model, making it suitable for research citation network. For each time slice, they calculate the topic respectively and then compare and analyze it with the topic computed on the last time slice to get the evolution situation of the topic. In addition, each constraint in this simple algorithm is removed respectively. For example, the

topic not only relies on the document of the present time slice, but also relates to the information of previous time slice. Therefore, the authors propose a topic inheritance model where they clearly point out how to apply references to topic evolution. In this model, the reference is interpreted as the succession of the topic and the time properties of the document is carefully considered, even the time sequence of documents in a time slice is to be measured, ultimately reflected as a partial order of a reference graph.

In reference [30], the research topic evolution graph is proposed, which includes the following elements: topic milestone paper, topic time strength and topic keyword.

(1) Topic milestone papers are the most representative of a topic in the process of understanding a topic.

(2) The topic time strength indicates the number of relatively related topic citations at different time, which mainly reveals the changes between current citations and previous citations and the life cycle of a topic.

(3) The topic keywords refer to the keywords that can accurately sum up the topic, which enables users to obtain a general understanding about the topic even before reading the relevant references and thus helps users to accurately locate papers they interest most in and they are supposed to read most in general references.

The biggest difference between this paper and previous research lies in that it not only considers the similarity between texts, but also takes into account the dependencies between cited papers. The authors verify that there is topic similarity between the papers citing the same paper, which reflects the topic similarity between papers more accurately than that of the texts.

Another feature of this paper lies in regarding each paper as a set of references, which then can be modeled with a topic generation model, with citations mainly represented by the latent topic variable. This is different from traditional approaches using the probabilistic topic model to discovery topic in the document. The generated topic of this article is based on a polynomial distribution of research papers, while the topic generated in traditional methods is the polynomial distribution of words. On this basis, taking into account the time factors of papers and references, we can get the exact topic evolution graph.

2. Other Topic Based Topic Evolution Models

Although the LDA is the most commonly-applied topic model approach on the issue of topic evolution in social networks, some other type of topic models are proposed when

taking into account the relevant background knowledge of topic research and the characteristics of data in social networks.

In the analysis of online social network community, an important task is to track the evolution situation of a topic in the community. But the existing methods consider only the emergence of the topics or the evolution of network structures, ignoring the interaction between text topics and network structures. Therefore, Lin et al. puts forward a Popular Event Tracking (PET) method in the reference [14]. This method takes into account the emergence of a user's new interest, information spread on the network structure, and the evolution situation of text topics. Gibbs Random Field is introduced to model the impact of historical status and dependency relationship on the graph, so that the topic model can be used to generate the corresponding keywords in the case of given interesting topics. Since Gibbs Random Field and the topic model are interacted, topic evolution becomes an optimization problem of a joint probability distribution including historical factors, text factors and structural factors. The authors also verify that the classic topic model is a special case of this model.

For the study of online social networks, especially the phenomenon of short text encountered during topic evolution in Twitter, the traditional text methods cannot solve this problem well due to the requirement of short time, a large amount of data to be processed and data sparse. Specifically, in the continuous generated data stream, we need to find tweets related to a preset topic. In this context, the incoming data stream is huge, the specified data related topics are very limited and it is required to complete the topic evolution analysis with both time and space constraints. On account of the requirement of about only one millisecond to process a tweet in practice, Lin et al. use a simple language model in the reference [16], particularly, using the label appeared in parts of tweets as an approximation to train the probability model, which can be applied to the continuous tweets in online social networks, with even most of those unlabeled tweets included. This paper adopts smoothing techniques to integrate timeliness and sparsity and also takes into account a variety of technologies for preserving historical information. Experiments validate that in this method the most appropriate smoothing technique is Stupid Backoff – the simplest smoothing technique. This paper shows that under the extreme circumstances of a large amount of data and the requirement of fast computing speed in online social networks and with consideration of equilibrium speed, availability, scalability and other demands for practical applications, the simplest way is the best way.

As for the cyclical phenomenon occurred in the topic evolution of social media (such as Flickr and Twitter), Yin et al. ^[31] in 2011 propose a Latent Periodic Topic Analysis (LPTA) based probability model. This model mainly considers addressing the following several issues:

(1) The existing work on the periodicity is generally concentrated in the time series database, and it is not suitable for text processing required in topic evolution analysis.

(2) Periodic word analysis can't satisfy the requirement of periodic topic analysis, because it is unlikely to see the same word again, but other words within the same topic.

(3) Due to the diversity of language, there are a lot of synonyms in text, which makes topic analysis a challengeable problem.

Therefore, the proposed LPTA method can be considered as a variant of latent topic model and the difference from traditional topic model lies in the time domain oriented period property. Specifically, it is to cluster the words of the same topic separated by about a time period. In other words, the problem is transformed into the estimation of the time period and the determination of the topic. Technically speaking, in this paper, the authors use the maximum likelihood probability approach to estimate the relevant parameters. The goal of LPTA is not only to identify latent topic space according to the data, but also to discover whether there is a periodicity in topic evolution.

11.3.3 Adjacent Time Slice Association Based Topic Evolution

In addition to the most simple and intuitive simple topic evolution model described in section 11.3.1 and the most commonly-used topic model introduced in section 11.3.2, there are other types of models in the study of topic evolution, while the difference between these models and the aforementioned two models lies in that these models mainly focus on the link between history time slice and present time slice and relevant background of topics.

In the social media stream, users have the following natural requirements for the topic evolution:

(1) In social media, once there emerges a new topic, we need to detect it in time.

(2) The evolution of the interesting topics can be tracked in real time.

(3) The information provided should be ensured in a proper range to avoid information overload.

Saha et al. ^[22] point out that new information appears all the time in social media, leading

to the fact that to send all topic information to users is impossible, which thus requires a model proposed to infer topics of users' interests according to the topic information of historical time slices. In this article, the authors propose a system based on the strict Rigorous Machine Learning and optimization strategy for the online analysis of social media stream. Specifically, it applies the effective topic-related model and Non-negative Matrix Factorization Techniques. Different from the existing work, this article maintains the time continuity of topic evolution at the process of new topic discovery.

Unlike the above-mentioned papers on topic evolution, Lin et al.^[15] also reveal the topic's Latent Diffusion Paths in social networks except the simple analysis of topic evolution. After comprehensively considering the text document, social influence and topic evolution, they turn the problem into a Joint Inference Problem. Specifically, the authors propose a hybrid model that on the one hand can generate topic related keywords based on topic evolution and diffusion, and on the other hand can adjust the diffusion process with Gaussian Markov Random Field using the social influence of users' personal layer. This paper believes that in social networks the information users acquire comes more from the social links rather than strangers. So the researchers believe the topic is actually spreading in the social network structure. Based on the proposed topic diffusion and evolution model, for one thing the latent spread path graph can be determined for further determining the source of a topic; and for another, the properties of the time change of the topic can be inferred to track the new development of a topic and understand the regularity of its changes.

11.4 Summary

With the advancement of computing technologies and the popularization of Internet, online social network has been developed dramatically today with an explosive growth of information. Therefore, the traditional knowledge acquisition approaches have become difficult to keep up with the pace of knowledge production in the era of big data, which leads to a demand for intelligent processing of information in the network age and automatic discovery and acquisition of knowledge. Among these, the social network contains a large amount of valuable information for its real reflection of people's life. Therefore, it is particularly important for text mining and analysis for social networks. Topic is important information concerned by social network users. In addition, detecting accurately the topic and tracking the evolution of topic have a great reference value in

monitoring and guiding public opinions, business decisions and other aspects. This chapter introduces theories and technologies related to topic discovery and topic evolution with the hope that readers can have a general understanding of the relevant fields.

Topic discovery methods mainly include topic model, the traditional vector space model based method and the term co-occurrence graph based method, where topic model is the current mainstream approach, whereas the above methods are all required to be suitably modified to suit the characteristics of text information in social networks so as to improve the analysis results. Research on topic evolution in social networks is relatively limited and the main method is still concentrated in the method of time slice segmentation, with the more complex considering the relationship between time slices, etc.

Overall, although the research on topic discovery and evolution has been carried out for many years, the initial research field is relatively narrow and the method used is not mature enough, which can't simply be applied directly to the current online social network environments. In recent years, the research on topic discovery and evolution has made considerable progress and development, but we believe there are still many challenging issues to be solved:

(1) The formal definition and presentation of the topic. Although many researchers have made relatively complete definitions and presentations of the topic, the topic itself involves some subjective factors and researches have not established a more systematic concept to describe it, which leads to more or less differences between researches of different fields of topic discovery and evolution and a barrier hindering more researchers from devoting to relevant studies.

(2) Online real-time processing of mass information. The world today is the era of information when timeliness of information has strategic significance. The requirements for timeliness of topic discovery and evolution in social networks has become higher and higher, so future research should focus much on topic discovery and evolution in massive online real-time information. Nowadays, Twitter, Weibo and other applications generate thousands of data per second, so how to adapt to such a rapid information update speed has become a problem that we must consider first in constructing related applications. In addition, we should try to study new information compressing representations and online processing algorithms to better meet the requirements of online social networks.

(3) Aggregation of multiple source information. With the rapid development of Internet, various social applications have sprung up and users no longer focus on a single information source. Aggregation for the multiple sources information to construct

multi-granularity and multi-level topic discovery and evolution application systems can better describe the topic and help users grasp the topic more intuitively and precisely.

References

- [1] Paige H. Adams, Craig H. Martell. Topic detection and extraction in chat [C]. 2008 IEEE International Conference on Semantic Computing, 2008.
- [2] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang. Topic Detection and Tracking Pilot Study Final Report [C]. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.
- [3] Hila Becker, Mor Naaman, Luis Gravano. Learning similarity metrics for event identification in social media [C]. Proceedings of the third ACM international conference on Web search and data mining, 2010.
- [4] Hila Becker, Mor Naaman, Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter [C]. ICWSM 11, 2011: 438-441.
- [5] David M. Blei, Andrew Y. Ng, Michael I. Jordan, John Lafferty. Latent Dirichlet allocation [J]. Journal of Machine Learning Research 3, 2003: 993-1022.
- [6] Levent Bolelli, Seyda Ertekin, C. Lee Giles. Topic and Trend Detection in Text Collections using Latent Dirichlet Allocation [C]. In ECIR'09, 2009.
- [7] Kai Cui, Bin Zhou, Yan Jia, Zheng Liang. LDA-based Model for Online Topic Evolution Mining [J]. Computer Science, 2011, 37(11): 156-159.
- [8] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by Latent Semantic Analysis [J]. JASIS, 1990, 41(6): 391-407.
- [9] Lei Deng, Zhaoyun Ding, Bingying Xu, Bin Zhou, Peng Zou. Using Social Intelligence for New Event Detection in Microblog Stream [C]. 2012 Second International Conference on Cloud and Green Computing (CGC), 2012: 434-439.
- [10] Lei Deng, Bingying Xu, Lumin Zhang, Yi Han, Bin Zhou, Peng Zou. Tracking the Evolution of Public Concerns in Social Media [C]. Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, 2013: 353-357.
- [11] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, C. Lee Giles. Detecting Topic Evolution in Scientific Literature: How Can Citations Help? [C]. CIKM'09, 2009.
- [12] Thomas Hofmann. Probabilistic Latent Semantic Indexing [C]. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.
- [13] Liangjie Hong, Brian D. Davison. Empirical Study of Topic Modeling in Twitter [C]. Proceedings of

- the First Workshop on Social Media Analytics, 2010: 80-88.
- [14] Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, Jiawei Han. PET: a Statistical Model for Popular Events Tracking in Social Communities [C]. KDD'10, 2010.
 - [15] Cindy Xide Lin, Qiaozhu Mei, Jiawei Han, Yunliang Jiang, Marina Danilevsky. The Joint Inference of Topic Diffusion and Evolution in Social Communities [C]. ICDM'11, 2011.
 - [16] Jimmy Lin, Rion Snow, William Morgan. Smoothing Techniques for Adaptive Online Language Models: Topic Tracking in Tweet Streams [C]. KDD'11, 2011.
 - [17] Yue Lu, Chengxiang Zhai. Opinion Integration through Semi-supervised Topic Modeling [C]. Proceedings of the 17th international conference on World Wide Web, 2008.
 - [18] Omid Madani, Jiye Yu. Discovery of Numerous Specific Topics via Term Co-occurrence Analysis [C]. Proceedings of the 19th ACM international conference on Information and knowledge management, 2010.
 - [19] Daniel Ramage, Susan Dumais, Dan Liebling. Characterizing Microblogs with Topic Models [C]. ICWSM, 2010.
 - [20] Daniel Ramage, David Hall, Ramesh Nallapati, Christopher D. Manning. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora [C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009.
 - [21] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, Padhraic Smyth. The author-topic model for authors and documents [C]. Proceedings of the 20th conference on Uncertainty in artificial intelligence, 2004.
 - [22] Ankan Saha, Vikas Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization [C]. Proceedings of the fifth ACM international conference on Web search and data mining, 2012.
 - [23] Gerard M Salton, Andrew Wong, Chungshu Yang. A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.
 - [24] Hassan Sayyadi, Matthew Hurst, Alexey Maykov. Event Detection and Tracking in Social Streams [C]. ICWSM, 2009.
 - [25] Hassan Sayyadi, Matthew Hurst and Alexey Maykov. Event Detection and Tracking in Social Streams [C]. ICWSM'09, 2009.
 - [26] Hassan Sayyadi, Louiqa Raschid. A Graph Analytical Approach for Topic Detection [J]. ACM Transactions on Internet Technology (TOIT), 2013, 13(2): 4.
 - [27] Jing Shi, Meng Fan, Wanlong Li. Topic Analysis Based on LDA Model [J]. Acta Automatica Sinica, 2009, 35(12): 1586-1592.
 - [28] Jintao Tang, Ting Wang, Qin Lu, Ji Wang, Wenjie Li. A wikipedia based semantic graph model for

- topic tracking in blogosphere [C]. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, 2011.
- [29] Ville H. Tuulos, Henry Tirri. Combining topic models and social networks for chat data mining [C]. Proceedings of the 2004 IEEE/WIC/ACM international Conference on Web intelligence, 2004.
- [30] Xiaolong Wang, Chengxiang Zhai and Dan Roth. Understanding Evolution of Research Themes: a Probabilistic Generative Model for Citations [C]. KDD'13, 2013.
- [31] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, Thomas Huang. LPTA: A probabilistic model for latent periodic topic analysis [C]. 2011 IEEE 11th International Conference on Data Mining (ICDM), 2011.
- [32] Wayne Zhao, J Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, Xiaoming Li, . Comparing twitter and traditional media using topic models [C]. Proceedings of the 33rd European conference on Advances in information retrieval, 2011: 338-349.

Appendix

1. Document 20140309 in Application Examples

We are working with anti-terrorism units, says Hisham

SEPANG: Malaysia has informed counter-terrorism agencies of various countries in light of several imposters found to have boarded flight MH370 that went missing over the South China Sea early Saturday.

Defence Minister Datuk Seri Hishammuddin Hussein (pic) said Malaysia would be working with intelligence agencies, including the Federal Bureau of Investigation (FBI), on the matter.

“If it is an international network, the Malaysian immigration alone will not be sufficient.”

“We have also informed the counter-terrorism units of all relevant countries.”

“At this point, we have not established if there was a security risk involved (and) we do not want to jump the gun,” Hishammuddin said when asked if there could be any hijack or terror elements in the disappearance of the MH370 flight.

On two impostors who boarded the flight using passports reported lost by an Italian and an Austrian, Hishammuddin said the authorities would screen the entire manifest of the flight.

2. Document 20140323 in Application Examples

MH370: Indian aircrafts join search

KUALA LUMPUR: Two long range maritime reconnaissance aircrafts from India have departed the Subang Airport on Sunday to join in the search for the missing Malaysia Airlines (MAS) MH370 aircraft in the southern Indian Ocean.

This was despite reported bad weather caused by tropical cyclone Gilia which had forced a number of other aircrafts scheduled to go on the mission to be cancelled.

The High Commission of India to Malaysia on Sunday said that both aircrafts were likely to encounter the cyclonic conditions enroute, but Captains of both aircrafts had instead “decided to skirt bad weather areas” to reach the search sectors.

The two aircrafts were the P8-I of the Indian Navy and the C-130J of the Indian Air Force. Both were expected to undertake a 10-hour mission in the area.

“Both aircrafts have long endurance capabilities coupled with state of the art electro optronic and infrared search and reconnaissance equipment on board,” said the high commission in a statement here.

The P8-I aircraft has the added advantage of on-board radars and specially designed search and rescue kits.

Previously, the Prime Minister of India had committed to assisting Malaysia and “render all possible assistance to Malaysia in locating the missing Malaysian Airlines flight”.

Both aircrafts arrived in Malaysia on March 21. After extensive briefings in Malaysia on Sunday, both Indian aircrafts took off to be part of the Australian-led search on Sunday morning.

India has been participating in the search and rescue operation beginning March 11 in the Andaman Sea and Bay of Bengal.

3. Document 20140325 in Application Examples

New statement from Malaysia Airlines: Tan Sri MdNorMdYusof, Chairman of Malaysia Airlines:

The painful reality is that the aircraft is now lost and that none of the passengers or crew on board survived.

This is a sad and tragic day for all of us at Malaysia Airlines. While not entirely unexpected after an intensive multi-national search across a 2.24 million square mile area,

this news is clearly devastating for the families of those on board. They have waited for over two weeks for even the smallest hope of positive news about their loved ones.

This has been an unprecedented event requiring an unprecedented response. The investigation still underway may yet prove to be even longer and more complex than it has been since March 8th. But we will continue to support the families—as we have done throughout. And to support the authorities as the search for definitive answers continues.

MAS Group CEO, Ahmad Jauhari Yahya, has said the comfort and support of families involved and support of the multi-national search effort continues to be the focus of the airline. In the last 72 hours, MAS has trained an additional 40 caregivers to ensure the families have access to round-the-clock support.

A short while ago Australian Defense Minister David Johnston said “to this point, no debris or anything recovered to identify the plane” He also said this is an extremely remote part of the world and “it’s a massive logistical exercise.” “We are not searching for a needle in a haystack. We are still trying to define where the haystack is,” he said.

4. Document 20140327 in Application Examples

Chinese celebs lash out at M’sia over MH370

A number of big-name celebrities in China, including award-winning actress Zhang Ziyi, have hit out at Malaysia’s handling of the search for Malaysia Airlines Flight MH370.

Several celebrities took to Weibo—a Shanghai-based Twitter-like microblogging service - to condemn Malaysia and urge a boycott of Malaysian products.

“The Malaysian government has offended the world! We are looking for the airplane but you are more concerned about timing.”

“Malaysian government, today you have done wrong. You are wrong failing to take responsibility. You are wrong for prioritising political manoeuvres instead of respecting life.”

“You are wrong for failing to respect the universal ... quest for truth,” wrote Zhang on March 25, 17 days MH370 went missing and a day after Prime Minister Najib Abdul Razak revealed that the plane went down in the Indian Ocean.

Among the 239 people on board MH370 were 152 Chinese nationals. One of them was Jo Kun, who was once Zhang’s fight choreographer.

Chen Kun

Actor Chen Kun (right), another famous Chinese celebrity, accused the Malaysian government and MAS of “clownish prevarication and lies”, also on March 25.

He added that he would boycott all Malaysian products and avoid coming to Malaysia

indefinitely.

Posts by Zhang, Chen and other Chinese celebrities have been widely shared online by Chinese netizens.

Fish Leong

Some Chinese netizens have also urged a boycott of Malaysian artistes such as Fish Leong (right), Gary Chaw, Lee Sinje and Ah Niu, who are popular for their music and movies.

Bahau-born Leong, who is an expectant mother, drew scorn from Microblogging users after uploading a photograph of three candles as a mark of respect for MH370 victims.

Numerous Chinese netizens responded by cursing her and her unborn child. The photograph has since been removed.

In an apparent attempt to stem the anger and distrust in China, Malaysian officials have also met the Chinese ambassador to Malaysia Huang Huikang to ask for the Chinese government to engage and help clarify the situation to the bereaved relatives and the public.

“Malaysia is working hard to try and make the briefings to the Chinese relatives in Beijing more productive,” read a statement sent out by the Transport Ministry today.

5. Document 20140405 in Application Examples

Chinese ship searching for missing Malaysia plane detects signal

BEIJING (Reuters) - A Chinese patrol ship searching for missing Malaysia Airlines flight MH370 detected a pulse signal with a frequency of 37.5 kHz per second in the south Indian Ocean on Saturday, state news agency Xinhua reported.

37.5 kHz per second is currently the international standard frequency for the underwater locator beacon on a plane’s “black box”.

A black box detector deployed by the vessel Haixun 01 picked up the signal at around 25 degrees south latitude and 101 degrees east longitude, Xinhua said. It has yet to be established whether it is related to the missing jet.

Xinhua also said a Chinese air force plane spotted a number of white floating objects in the search area.

(Reporting by Benjamin Kang Lim; editing by Andrew Roche)

Algorithms of Influence Maximization

12.1 Introduction

Social network is playing a fundamental role in the spread and diffusion of information, viewpoints and innovation. With the purpose of discovering the most information-spread influential node set in social networks, the problem of Influence Maximization is one of the key issues being researched in the field of social network information spread, and is widely applied to many important occasions, such as marketing, advertising publication, early warning of public sentiment, water quality monitoring, disease surveillance and control and so on, so it is of high research value and application value. For instance, in the word-of-mouth marketing and advertisement publication which are based on social networks, what kind of users will be used for the product and advertising promotion so as to maximize, through information and influence spread in the social network, the promotion profits of the brand and the spread scope of advertisement^[1]. The solution of Influence Maximization will directly affect the formulation and deployment of such application strategies as marketing or the like, and has an important influence on the effectiveness and scalability and so on of the system.

With the development of social network technology, the current social network is larger and larger, specifically reflected by numerous nodes and complex correlation between nodes; meanwhile, the network is more and more dynamic, specifically reflected by the frequent variation of the number of the nodes and the correlation between nodes, high randomness and unpredictability. These characteristics of social network directly lead

to a large amount of calculation and a long run time for seeking the most influential node in the network. Since the requirement from practical applications for the execution time of algorithm is more and more strict, it is urgently needed to deeply study the efficient processing technology of Influence Maximization in the environment of large-scale social networks.

The contents of this chapter is arranged in a way as follows: in section 12.2, Influence Spread models involved in the study of Influence Maximization will be introduced, and a formal definition of the problem of Influence Maximization will be given. In section 12.3, three basic metric for measuring the Influence Maximization Algorithms will be introduced: the run time, the algorithm precision, and scalability. In section 12.4, the algorithms of Influence Maximization are classified as Greedy Algorithms and Heuristic Algorithms. In section 12.5, Greedy Algorithms of Influence Maximization, including: BasicGreedy, CELF and MixGreedy, and other Greedy Algorithms will be introduced, and their advantages and disadvantages will be concluded. In section 12.6, several heuristic algorithms of Influence Maximization will be introduced, and their advantages and disadvantages will be concluded. In section 12.7, related researches on the extension and deformation of Influence Maximization will be elaborated.

12.2 Basic Concepts and Theory Basis

The modeling basis of Influence Maximization is the topology of social networks and corresponding models of Influence Spread. Therefore, Influence Spread models involved in the researches of Influence Maximization will firstly be introduced in this section, and the formal definition of Influence Maximization will be given.

Concept 1: Influence Spread models

The Influence Spread models defines the methods and mechanisms of Influence Spread in social networks, and is the basis for studying the Influence Maximization problem. Different from the afore-mentioned topology of social network, each node in the social network has two states, which are respectively 0 and 1, wherein 0 represents an inactive state while 1 represents an active state. After a node v_i switches from an inactive state to an active state, this node v_i will try to influence other inactive neighboring nodes. If the activation succeeds, then the neighbors will switch from the inactive state into the active state. As shown in Figure 12-1, the node a is in the active state at the beginning, so it is capable of trying to influence other neighboring inactive nodes b , c and e . There

may be two situations at this moment: one situation is that the node a fails to activate the node c , then the node c will stay in the inactive state; the other situation is that the node a activates the node b successfully, so the node b switches from the inactive state into the active state, and is currently capable of influencing other nodes, for instance, the node b may activate the node d . The process of influencing a node to switch from the inactive state into the active state is called as the spread of influence. It should be noted that the whole spread process is irreversible, namely, one node may be influenced to switch from the inactive state into the active state, but not vice versa.

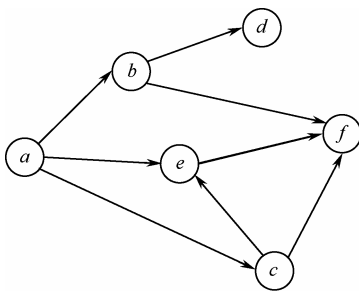


Figure 12-1 Diagram of influence spread

In-depth researches on the Influence Spread models have been made in the current academic community, wherein the Independent Cascade Model^[2] and the Linear Threshold Model^[3] are currently the two most widely researched Influence Maximization models.

The Independent Cascade Model is based on the probabilities, wherein each node, after switched into the active state itself, will try to active its subsequent node at a certain probability, and the behaviors for plural active nodes to try to activate the same neighboring node are independent from each other, so the model is named as the Independent Cascade Model. The Linear Threshold Model is based on thresholds, the behaviors for plural active nodes to try to activate the same subsequent node are dependent, and whether the influence is successful depends on whether the sum of the weights of the influences on the same subsequent node by all the active nodes surpasses the threshold of the subsequent node. The Independent Cascade Model and the Linear Threshold Model elaborates the process of Influence Spread respectively from the angle of probability and threshold. Please refer to Chapter 10 for details, and pleonasm will be omitted here.

Besides the above models, there are also some other Influence Spread models in the research of Influence Maximization. Brief description is as follows.

(1) Triggering Model^[1]. The Triggering Model was brought up by David Kempe et al. in 2003. In Triggering Model, each node v is corresponding to a triggering node set T_v which defines the nodes capable of triggering the node v to switch from the inactive state into the active state; namely, at the time point of t , a non-activated node v will be activated if and only if the precursor u of at least one node v is in the active state at the time point of $t-1$, and $u \in T_v$. Take Fig. 12-1 as an example, suppose that $T_e = \{a, c\}$,

$T_f = \{b, e\}$, and the node c is in the active state when $t = T_0$. Then, when $t = T_1$, c tries to influence nodes e and f . Since $c \in T_e$ but $c \notin T_f$, c can successfully activate the node e , while f is still in a non-activated state. When $t = T_2$, the newly activated node e also tries to influence its neighboring node f . since $e \in T_f$ at this moment, the node e succeeds in activating the node f .

(2) Weighted Cascade Model^[4] is an unique Independent Cascade Model. The difference between the Weighted Cascade Model and the Independent Cascade Model lies in that, in the Weighted Cascade Model, the probability for the node v to successfully activate the subsequent node w is the reciprocal of the in-degree of the node w , namely, $p(v, w) = 1/d_w$, wherein d_w is the in-degree of the node w . For instance, in Figure 12-1, the probability for the node e to successfully activate the node f is $p(e, f) = \frac{1}{3}$.

(3) Voter Model^[5] is a probability model which is widely applied to Statistical Physics and Particle Systems, and was presented by Peter Clifford and Aidan Sudbury. In the Voter Model, each node randomly selects one node from its precursor node set in each step, and takes the state of the selected node as its own state. Take Figure 12-1 as an example again, suppose that only the node c of all nodes is in the active state when $t = T_0$, while all the other nodes are inactive. Then, when $t = T_1$, the node e randomly selects one node from its precursor node set $\{a, c\}$; if the node a is selected, then the node e is still inactive at the time point of T_1 ; otherwise, if the node c is selected, then e will switch from the inactive state into the active state at the time point of T_1 . It should be noted that, the states of nodes in the Voter Model can either switch from the inactive state into the active state, or switch from the active state into the inactive state, so the Voter Model is more suitable for modeling those occasions allowing for the nodes to change their viewpoints, e.g., the public in democratic elections may change their votes due to the influence of the votees and other people.

Concept 2: Influence Maximization

Based on the given Influence Spread Model, the purpose of Influence Maximization is to find K nodes having the largest ultimate influence spread in social networks so that maximum nodes will be ultimately influenced through the Influence Spread in social networks with these K nodes as the initial active node set. For formal description of Influence Maximization, when the symbol $\sigma(S)$ is used for representing the number of the nodes ultimately activated by the initial seed set S after the process of influence spread. On such a basis, the formal description of

Influence Maximization is as follows:

Given: a social network $G=(V,E,W)$, a given Influence Spread Model, and a positive integer K ;

Goal: to select an initial active node set S from the network G so that $\sigma(S)$ is the largest;

Constraints: $S \subset V$ and $|S|=K$.

12.3 Metrics of Influence Maximization

Practical applications of Influence Maximization have made lots of requirements for the solution of the problem, wherein the run time, the algorithm precision and the scalability are key factors need to be considered in solving the problem of Influence Maximization in the current environment of large-scale social networks. Nodes of large-scale social networks are numerous, and the connections are tight and complicated, as a result, the run time is long, and it is hard to deal with the problem. As the requirements for algorithm efficiency from modern applications are more and more strict, the run time becomes the primary standard for measuring Influence Maximization Algorithms; moreover, in the problem of Influence Maximization, higher precision means a larger ultimate influence range, so algorithm precision is also a key factor needs to be considered in the design of Influence Maximization Algorithms. Meanwhile, the constantly increasing scale of social networks and mass data have brought serious challenges to the scalability of Influence Maximization Algorithms. In a word, the run time, the algorithm precision and the scalability are three essential standards for measuring the quality of Influence Maximization Algorithms.

1. Run Time

The computation time of Influence Maximization Algorithm is significant for the policy formulation and deployment of many important applications, so this type of applications are very sensitive to the computation time of algorithms, and have very high demands on algorithm efficiency. For example, in such application occasions as water quality monitoring and disease surveillance and control, delayed discoveries will cause severe dangerous of widespread pollution of water area and massive disease outbreak. Therefore, the run time is one of the core optimal objects for designing Influence Maximization Algorithms, and the possibly shortest run time must serve as

the basic starting point for designing the solution scheme of Influence Maximization. However, traditional Greedy Algorithm of Influence Maximization calculates the influence range of a given node set through a large amount of repeated Monte-Carlo simulations, which causes a rather long run-time. Particularly, in face of current large-scale social networks, existing algorithms cannot meet the requirements of applications for algorithm efficiency. Therefore, the run time of algorithm is one of the key metrics of Influence Maximization.

2. Algorithm Precision

As for Influence Maximization, algorithm precision refers to the number of nodes ultimately influenced by the seed set selected by the Influence Maximization algorithm after the process of influence spread. In practical application occasions of Influence Maximization, many applications require a possibly largest ultimate influence range. Applications of this type are represented by marketing and advertisement publication or the like. In the above two application, a larger ultimate Influence Maximization range indicates better promotion benefits of the product and more commercial profits. Therefore, the exploration of high-precision computation algorithm is also a key issue in the Influence Maximization researches. Influence Maximization researches in recent years prove that it is a NP-Hard problem to find the most influential K nodes^[1]. Traditional sorting algorithm, PageRank and other methods pay no attention to the characteristics of Influence Spread, so the algorithm precision is too low to solve the Influence Maximization problem. As a result, high precision is also a goal sought by Influence Maximization Algorithms.

3. Scalability

Scalability is an important metric for practical application of Influence Maximization. Due to the complicated algorithm and the long run time, the current solution algorithm can be applied only to the social network of with less than a million nodes. Faced with the large-scale social networks, Influence Maximization Algorithms with fine scalability must be designed for handling the severe challenge brought by the mass data of social network.

12.4 Classification of Influence Maximization Algorithms

In recent years, Influence Maximization Algorithms have caused extensive researches and attention from academic circles. Since Influence Maximization is proved to be a NP-Hard problem^[1], the research Influence Maximization can mainly be divided into two directions:

(1) Greedy Algorithm. Researches on Greedy Algorithm are basically on the basis of Hill-Climbing Greedy Algorithm^[1], wherein one node capable of providing the maximum influence value is selected at each step, and a locally optimal solution will be used for approximating the globally optimal solution. The advantage of Greedy Algorithm is its comparatively high precision, which can reach an approximate optimal of $1 - 1/e - \varepsilon$. However, Greedy Algorithm has the serious problem in efficiency, namely, the high algorithm complexity and the long run time. As a result, it can hardly be applied to large-scale social networks. Numerous studies and related optimization have been made specific to the efficiency of Greedy Algorithm, and this problem is still a hot topic in current research.

(2) Heuristic Algorithm. Different from Greedy Algorithm, Heuristic Algorithm selects the most influential node according to a designed heuristic strategy, and does not have to calculate the precise influence value of the node, so the run time of Heuristic Algorithm is short, and the efficiency is high. However, its algorithm precision is too low to compare with Greedy Algorithm.

12.5 Greedy Algorithm of Influence Maximization

In order to enable the readers to better understand the Greedy Algorithm of Influence Maximization, first, the following two key concepts will be introduced in this section: marginal profit and sub-modular function; then, typical Greedy Algorithm of Influence Maximization such as BasicGreedy, CELF and MixGreedy will be introduced in detail, and advantages and disadvantages of Greedy Algorithm will be concluded.

12.5.1 Basic Concepts of Greedy Algorithm

Concept 1: Marginal profit

For Influence Maximization, the marginal profit of the influence value function $\sigma(\cdot)$ refers to the growth of ultimate influence value that can be brought by extra addition of a node v_i as the initial active node on the basis of the current active node set S . That is

$$\sigma_{v_i}(S) = \sigma(S \cup \{v_i\}) - \sigma(S)$$

Concept 2: Sub-modular function

For an arbitrate function $f(\cdot)$ which maps the subset of a finite set U to be non-negative real numbers, if the function $f(\cdot)$ satisfies decreasing profits, then the function $f(\cdot)$ is a sub-modular function. The decreasing profits here means that the marginal profit brought to the set S by the addition of any element v_i is not lower than the marginal profit brought to the superset $T \supseteq S$ of S by the addition of the element v_i , the formal description is as follows:

$$f_{v_i}(S) \geq f_{v_i}(T)$$

or

$$f(S \cup \{v_i\}) - f(S) \geq f(T \cup \{v_i\}) - f(T)$$

Basic theory 1:

If the function $f(\cdot)$ is a sub-modular function as well as a monotonic function ($f(S \cup \{v_i\}) \geq f(S)$ satisfies all sets S and all elements), when it is tried to locate the element set S with a size of K so that $f(S)$ is maximum, Hill-Climbing Greedy Algorithm can be used to obtain the approximate optimal solution of $1 - 1/e - \varepsilon$ ^[6,7], wherein e is the base of a natural logarithm, and ε can be any positive real number.

12.5.2 Basic Greedy Algorithm^[1]

Pedro Domingos and Matt Richardson^[8] studied Influence Maximum as an algorithm problem for the first time. The earliest method is to regard the market as a social network, model the individual purchasing behavior and the overall earning promotion after marketing as a model of Markov Random Field, and bring up Single Pass Algorithm and Greedy Search Algorithm so as to gain an approximate solution.

On such a basis, David Kempe et al.^[1] firstly refined this problem to be a discrete optimization problem, that is to find K nodes capable of maximizing the ultimate influence range according to a given spread model. Kempe and et al. proved that this optimization problem was a NP-Hard problem in both the Independent Cascade Model and

the Linear Threshold Model. Later, the author proved that the influence value function $\sigma(\cdot)$ satisfied the sub-modular character and the monotonic character in both Influence Spread Models, and thus put forward a Greedy Hill-Climbing Approximate Algorithm BasicGreedy, which is capable of ensuring the approximate optimal of $1 - 1/e - \varepsilon$.

The Greedy Hill-Climbing Algorithm presented by Kempe and et al. is shown as the Algorithm 12-1. The algorithm, starts when S is a null set (Line 1), and later executes K rounds (Line 2), a node v capable of providing the maximum marginal profit will be selected in each round (Round 10), and then is added into the initial node set S (Line 11). In order to calculate the marginal profit s_v of each node in graph G (Line 3), Kempe and et al. designed to calculate, through R rounds of stimulation (Lines 5~7), the number of nodes which can be ultimately influenced with the set $S \cup \{v\}$ as the initial active node in each round, and finally seek the average value (Line 8) and selects the node having the largest marginal profit to join the set S .

Algorithm 12-1 Greedy Hill-Climbing Approximation Algorithm Basic Greedy
<p>Known: Social Network $G = (V, E, W)$, Parameter K</p> <p>Seek: the most influential node set S with a size of K</p> <p>1: Initialize $S = \emptyset$;</p> <p>2: for $i = 1$ to K do</p> <p>3: for any node v in the set $V \setminus S$ do</p> <p>4: $s_v = 0$;</p> <p>5: for $i = 1$ to R do</p> <p>6: $s_v = s_v + \text{RanCas}(S \cup \{v\})$;</p> <p>7: end for</p> <p>8: $s_v = s_v / R$;</p> <p>9: end for</p> <p>10: $v = \text{argmax}_{v \in (V \setminus S)} \{s_v\}$;</p> <p>11: $S = S \cup \{v\}$;</p> <p>12: end for</p>

However, the complexity of BasicGreedy Algorithm is as high as $O(Knm)$, wherein n and m are respectively the number of nodes and the number of edges in graph G , R is the number of stimulation, which generally selects a value of 20000. therefore, BasicGreedy Algorithm has a rather long execution time, and cannot be applied to large-scale social networks. There are two main reasons for its time-consuming: ① The

marginal profits of all nodes need to be calculated in each round of the algorithm; ② R stimulations are needed for calculating the marginal profit of each node. Therefore, although Greedy Hill-Climbing Approximation Algorithm ensures fine precision, its low calculation efficiency demands prompt solution. Basically all of the subsequent Greedy Algorithm researches aim at improving its efficiency.

12.5.3 CELF Algorithm^[9]

In 2007, Jure Leskovec et al.^[9] proposed CELF (Cost-Effective Lazy Forward) which is an optimization of BasicGreedy. Since the Influence Value Function $\sigma(\cdot)$ satisfies the sub-modular character, the Influence Value marginal profit brought by a random node v has to be smaller following the growth of the initial active set S . On such a basis, CELF Algorithm does not have to calculate in each round the Influence Value marginal profits of all nodes like BasicGreedy Algorithm does. If the Influence Value marginal profit of the node u prior to this round is smaller than that of the node v in the current round, then the Influence Value marginal profit of the node u in the current round is bound to be smaller than that of the node v , so it is impossible for the node u to be the node having the largest marginal profit in the current round, and there is no need to calculate its Influence Value marginal profit in the current round. Exactly by using the sub-modular character of the Influence Maximization Objective Function, CELF Algorithm greatly reduces the number of calculations of the Influence Value marginal profits of nodes in each round, and narrows the selection range of nodes, thereby lowering the overall calculation complexity. Experiment results show that the precision of CELF Algorithm is basically the same as that of BasicGreedy Algorithm, while its calculation efficiency is far higher than that of the BasicGreedy Algorithm, and reaches an acceleration which is as high as 700 times. Even so, it still costs hours for CELF Algorithm to seek 50 most influential nodes in a data set having 37,000 nodes, and its efficiency can hardly satisfy the demand of short run time by current social networks.

12.5.4 Mix Greedy Algorithm^[4]

Wei Chen et al.^[4] presented novel Greedy Optimization Algorithm, i.e. NewGreedy and MixGreedy. In the original BasicGreedy Algorithm, R stimulations are needed for calculating Influence Value marginal profits of each node, so there will be altogether nR stimulations for all the n nodes in the network, which causes a large amount of calculation.

NewGreedy Algorithm make improvement right on such a basis, and enhances the algorithm efficiency. The core thought of NewGreedy Algorithm lies in the calculation of the Influence Value marginal profits for all nodes in each stimulation, so NewGreedy Algorithm reduces the nR stimulations of BasicGreedy Algorithm to R . Specifically, in each stimulation, all the edges failed to be influenced will be removed from the original network, and a Network Spread Map will be obtained; then Breadth First Search (BFS) will be conducted for each node in the Network Spread Map and obtain the influence value for each node. Since it is time-consuming to perform Breadth First Search for each node, a random algorithm, which was proposed by Edith Cohen and et al., is adopted in the NewGreedy Algorithm for estimating the possible number of nodes in the Network Spread Map. Due to the Cohen Random Algorithm, on the one hand, the complexity of NewGreedy Algorithm is sharply reduced from $O(KnRm)$ for BasicGreedy to $O(KnTm)$, wherein T is the iteration times of Cohen Random Algorithm, and is far less than n ; but, on the other hand, a method of estimation is used in Cohen Random Algorithm, so it is impossible to obtain the accurate influence value of nodes, which correspondingly lowers its precision. The design of NewGreedy Algorithm is shown in Algorithm 12-2.

Algorithm 12-2 NewGreedy Algorithm
<p>Known: Social Network $G = (V, E, W)$, Parameter K</p> <p>Seek: the most influential node set S with a size of K</p> <p>1: Initialize the set to be a null set $S = \emptyset$;</p> <p>2: for $i = 1$ to K do</p> <p>3: Provide the influence value s_v of all nodes in Fig. G to be 0;</p> <p>4: for $j = 1$ to R do</p> <p>5: Remove the edges failed to be influenced in Fig. G according to IC Model so as to get the Fig. G';</p> <p>6: for any node v in Fig. G do</p> <p>7: Calculate the Influence Value marginal profit $MG(G', v)$ for node v;</p> <p>8: $s_v = s_v + MG(G', v)$;</p> <p>9: end for</p> <p>10: end for</p> <p>11: $v_{\max} = \operatorname{argmax}_{v \in (V \setminus S)} s_v / R$;</p> <p>12: $S = S \cup \{v_{\max}\}$;</p> <p>13: end for</p>

MixGreedy Algorithm is a combination of NewGreedy Algorithm and CELF Algorithm. In the first round of CELF Algorithm, the Influence Value marginal profits of all nodes need to be calculated, so the amount of calculation is rather large; however, thanks to the sub-modular character, it is not necessary to calculate the Influence Value marginal profits of a part of nodes after the first round, and the calculation amount is sharply reduced. For NewGreedy Algorithm, R simulations are needed in each round so as to calculate the Influence Value marginal profit for each node. The advantages of NewGreedy Algorithm and CELF Algorithm are orthogonal and are not in conflict, so the MixGreedy Algorithm takes the advantages of these two algorithms, that is to use NewGreedy Algorithm in the first round and use CELF Algorithm in succeeding rounds so as to reduce the calculation amount, thereby further reducing the overall algorithm complexity. The experiment results show that the NewGreedy Algorithm and the MixGreedy Algorithm can significantly accelerate the discovery of most influential users in social networks, and meanwhile ensure a precision which is basically the same as that of BasicGreedy.

12.5.5 Other Greedy Algorithms

In 2010, Yu Wang and others^[10] put forward a solution algorithm CGA which is based on the concept of community. Social networks demonstrate good community characteristics, namely, the interaction between community members is close, so the probability of being influenced by each other is rather high; on the contrary, connections between members of different communities is relatively few, so the probability of interaction is comparatively low. Exactly based on this community characteristic, Yu Wang and et al. presented the CGA Algorithm so as to approximate the most influential user in the global network by using the most influential user inside of the community, thereby reducing the calculation complexity. The execution of CGA Algorithm is divided into two phases: in the first phase, specific to the problem that the information spread factor is not taken into consideration in the current community division algorithm, the author designs a new combination entropy division algorithm on the basis of Influence Spread; in the second phase, CGA Algorithm selects the most influential user from the divided community by using the method of dynamic programming. Suppose that $k-1$ most influential nodes have been obtained in preceding $k-1$ rounds, each community will respectively serve as the object of study in Round k , and MixGreedy Algorithm will be

used in each community so as to select the most influential node, which will be selected as the globally most influential node in Round k . Through the experiments on mobile social network, the author proved that the operating speed of CGA Algorithm was significantly enhanced compared with MixGreedy Algorithm. However, the enhancement of speed is at the cost of precision, because CGA Algorithm approximates the global influence of a node by using its influence inside the community, which lowers the precision.

Amit Goyal and et al.^[11] thoroughly analyzed CELF Algorithm, and presented the CELF++ Algorithm, which is an optimization method specific to CELF Algorithm. CELF++ Algorithm once again uses the sub-modular character of the Influence Value Function $\sigma(\cdot)$, and respectively records in the current iteration for all nodes the most influential node ID: $\text{prev}_{\text{best}}$ after this node calculation. If the $\text{prev}_{\text{best}}$ node of the node v_i is selected as the most influential node in the current round after the iteration in current round, then the influence value of the node v_i does not need to be calculated in the iteration of the next round, thereby avoiding numerous recalculation of influence values existing in CELF Algorithm. The author proved by experiments that CELF++ Algorithm could reduce 35%~55% of the run time compared with CELF Algorithm.

On the basis of MixGreedy Alorithm, Xiaodong Liu and et al.^[12] deeply analyzed the layer dependency and parallelizability of the nodes in the social network, designs, by the transformation of a directed acyclic graph and the bottom-up layer-by-layer scanning, an Influence Maximization Algorithm BUTA having a high parallelizability so as to efficiently concurrently calculate the influence values of all nodes in the social networks. Later, a parallel computing system of CPU+GPU was taken as a representative of the existing heterogeneous-parallel computing frame, BUTA Algorithm was mapped onto the heterogeneous-parallel frame of CPU+GPU, and an IMGPU frame was proposed. In order to better make BUTA Algorithm adaptive to GPU hardware frame and programming models, the author provided the following three optimization methods: K -layer merging, data recombination and memory access merging, in order to reduce the number of branches and memory access and to improve parallelism. At last, a large amount of experiments were intensively performed in the real social networks, and the experiment results showed that the execution speed of BUTA Algorithm and IMGPU was notably improved relative to the aforementioned MixGreedy Algorithm.

12.5.6 Summary of Greedy Algorithms

To sum up, Greedy Hill-Climbing Approximation Algorithm laid a foundation for the Influence Maximization Approximation Algorithm. Although Greedy Hill-Climbing Algorithm ensures a high solving precision, it has a high complexity and a large calculating amount, resulting in a rather long run time. Numerous follow-up researches have been done to optimize this efficiency problem, and have achieved notable effect of acceleration, but still fail to meet the demand of high algorithm efficiency. Particularly, facing the current large-scale social networks, it is still the core object of current researches to design more efficient Influence Maximization Algorithm.

12.6 Heuristic Algorithms of Influence Maximization

Notable acceleration has been made for subsequently improving and optimizing Greedy Algorithms; however, due to the high complexity of Greedy Algorithms, the run time after optimization still cannot meet the requirement of short run time by current large-scale social networks. Meanwhile, in the pursuit of higher algorithm efficiency, many excellent heuristic algorithms have been proposed for shortening the run time of Influence Maximization. Researches of existing heuristic algorithms will be introduced in this chapter.

12.6.1 Degree Discount Heuristic^[4]

The most basic heuristic algorithms are Random, Degree and Centrality Heuristics proposed by David Kempe and et al.^[1]. Random Heuristic only randomly selects K nodes from all of the node sets V in the target social network G , and considers nothing about such factors as Influence Degree and Influence Spread. In comparison, Degree Heuristic and Centrality Heuristic are better, and both identify the most influential node in the network according to some network topological characteristics of the nodes. Degree Heuristic considers the concept of sociology, that is, measuring the node influence spread according to its degree. So Degree Heuristic ranks all nodes in the network according to their degrees, and selects K nodes having the largest degree. Similar to Degree Heuristic, Centrality Heuristic believes that the node, the average distance between which and other

nodes in the network is the smallest, has larger probability to influence other nodes, so it ranks the nodes according to the average distance between the node and other nodes in the network, and selects K nodes having the smallest average distance. Apparently, the design ideas of the above three basic heuristics are simple, so the execution time is as short as merely several seconds, or even several milliseconds. However, since they considers nothing about such factors as the actual Influence Value of nodes and Influence Spread, their algorithm precisions are pretty poor.

Based on the Degree Heuristic, Wei Chen and et al.^[4] put forward DegreeDiscount in 2009, which is a heuristic algorithm specific to Independent Cascade Model. The core thought of this heuristic is as follows: if a node u existing in the neighboring nodes of the node v is selected as an initial active node, the degree of the node v needs to be quantified and discounted due to the overlap existing between the two nodes. Details about the discounting method are shown in Algorithm 12-3. Experiment results show that the algorithm precision of DegreeDiscount Heuristic is much higher than that of Degree Heuristic, but still cannot be compared with the afore-mentioned Greedy Algorithms.

Algorithm 12-3 Degree Discount Algorithm

Known: Social Network $G = (V, E, W)$, Parameter K
 Seek: The most influential node set S with a size of K
 1: Initialize the set to be a null set $S = \emptyset$;
 2: **for** any node v in Fig. G **do**
 3: $dd_v = d_v$;
 4: $t_v = 0$;
 5: **end for**
 6: **for** $i = 1$ to K **do**
 7: $u = \operatorname{argmax}_v \{dd_v | v \in V \setminus S\}$;
 8: $S = S \cup \{u\}$;
 9: **for** any neighbor $v \in V \setminus S$ of the node u **do**
 10: $t_v = t_v + 1$;
 11: $dd_v = d_v - 2t_v - (d_v - t_v) \cdot t_v \cdot p$;
 12: **end for**
 13: **end for**

12.6.2 PMIA Heuristic^[13]

Specific to the Independent Cascade Model, Wei Chen and et al.^[13] put forward a new heuristic algorithm PMIA in 2010. Firstly, the author proved that the calculation of the Influence Value of the given initial active set in the Independent Cascade Model was a #P-Hard problem, then the author provided a heuristic algorithm PMIA which is based on local influence. PMIA has high efficiency and good scalability, because PMIA approximates the Global Influence Value of the node by using its Influence Value in its peripheral local area, constructs the Maximum Influence Arborescence (MIA) Model through the Maximum Influence Path, and compromises between the algorithm execution efficiency and the algorithm precision through the regulation of the MIA size. The author proves that the Influence Function is still in compliance with the sub-modular characteristics in MIA Model, so the Greedy Algorithms can reach an approximate optimal of $1 - 1/e - \varepsilon$. For higher execution efficiency, the author provided a heuristic PMIA, which is on the basis of MIA Model. PMIA merely needs to calculate the Influence Value of nodes in local area, and update the Influence Values of locally relevant nodes, so the calculation efficiency is higher. However, although PMIA heuristic improved the efficiency by means of local approximate optimal, but its precision is inevitably lost, which result in a over-low algorithm precision.

12.6.3 LDAG Heuristic^[14]

Also in 2010, Wei Chen and et al. presented for the first time the LDAG Heuristic specific to the Linear Threshold Models. The author first proved that the calculation of the Influence Value of the given initial active set in the Linear Threshold Models was a #P-Hard problem. The author discovered that the Influence Value of a node could be rapidly obtained in the Directed Acyclic Graph (DAG), so the author provided the efficient LDAG Heuristic which is based on DAG. This Heuristic is based on the principle of locality, establishes a local DAG for each node in the social network by removing part of the sides having over-low weight, then calculates in the constructed DAG graph the Influence Value of each node, and selects the node with the maximum Influence Value as the algorithm result. The experiments prove that LDAG Heuristic can remarkably accelerate the resolving of the Influence Maximization problem in the Linear Threshold Models. However, similar

to PMIA Heuristic, the speed improvement of LDAG Heuristic is also at the cost of algorithm precision.

12.6.4 Other Heuristics

In 2011, Amit Goyal et al.^[15] deeply analyzed LDAG Heuristic, and discovered the following disadvantages: ① Greedy strategy is used for constructing the Directed Acyclic Graph, which reduces algorithm precision; ② LDAG Heuristic merely takes into consideration the Influence Spread within LDAG, while many other Influence Spread paths exist in reality, and the Influence Spread via these paths are neglected by LDAG; ③ All of the Directed Acyclic Graphs are needed to be stored, so a large memory space is taken. With respect to these problems, the author designed the SIMPATH Heuristic, which can accurately estimate the Influence Value of the nodes by counting the number of paths starting from the seed node. Besides, the author also provided the Vertex Cover Optimization method and the Look Ahead Optimization method. The Vertex Cover Optimization method is configured to reduce the times of Influence Value Calculation in the first round of iteration, so as to reduce the algorithm complexity and shorten the algorithm execution time of the first round. Later, the Look Ahead Optimization method further reduces, via the parameter, the algorithm execution time in the process of Influence Value Calculation in the follow-up rounds. Experiments on real data sets prove that SIMPATH Heuristic is better than LDAG and other heuristics in algorithm execution time, algorithm precision, and memory utilization and other aspects.

Kyomin Jung et al.^[16] designed a new IRIE Heuristic in 2012 on the basis of Independent Cascade Models. Traditional heuristics and PMIA Heuristic obtains the Influence Value of nodes through rounds of simulation or by means of Local Influence Value Calculation, and thereby selecting the node having the maximum Influence Value. However, for large-scale social networks, it is rather time-consuming to calculate the Influence Value of all nodes. Therefore, IRIE Algorithm is novel for that IRIE does not need to calculate the Influence Value for each node; instead, it is based on the method of Belief Propagation, ranks the Influence Values of global node merely through rather few rounds of iteration, and then selects the top-ranked node as the most influential node. Moreover, IRIE is integrated with the method of Influence Estimation, estimates the influence of the most influential node on other nodes after each round of ranking, and then regulates next round of Influence Ranking according to the results. IRIE combines the

method of Influence Ranking and the method of Influence Estimation, and thus is averagely two orders of magnitude faster than Independent Cascade Model Heuristic PMIA, and is equivalently precise as PMIA.

Furthermore, researches in Literature [17] indicate that it is not necessary to precisely calculate the Influence Value of each node in social networks, and relative ranking according to node Influence Value will be enough. In addition, the distribution of social network nodes is subject to certain rules, so the nodes can be randomly sampled according to Monte Carlo Theory, and a distribution of overall sample can be approximated according to the distribution of the small sample, so as to approximate and estimate the node Influence Value, thereby decreasing the amount of calculation and improving algorithm execution efficiency. The author designed a supervised sampling method ESMCE based on power law index. Through deep analysis of node distribution features in social networks, ESMCE Algorithm determines the number of nodes in the initial sample according to power law index of the given social network; in order to minimize the number of the sample nodes and the number of sample rounds, ESMCE Algorithm proposed a method of forecasting the number of follow-up sample nodes based on a Grey Forecasting Model, which gradually refines the precision by means of iteration sampling till the error satisfies the predetermined requirement.

12.6.5 Summary of Heuristic Algorithms

The efficiency of Influence Maximization has been well improved through the afore-mentioned numerous researches on heuristic algorithms. These heuristic algorithms effectively shortens the algorithm execution time, but also causes server damage to the precision. The afore-mentioned heuristic algorithms select the most influential nodes by approximating or estimating the node Influence Value, or even without calculating the node Influence Value, so they have low complexity and short run time. But their precision cannot be guaranteed, and thus cannot be compared to the afore-mentioned Greedy Algorithms.

12.7 Extension and Deformation of Influence Maximization

As researches of Influence Maximization from academic circles go increasingly deeper, the Influence Maximization problem itself is continuously extended, deformed and

expanded, and is constantly used for solving more problems in social networks and other fields. Related researches about the extension and deformation of Influence Maximization will be elaborated in this section.

12.7.1 Extension of Influence Maximization

The subject of Influence Spread in basic Influence Maximization problem is a single subject, for instance, one certain kind of commodity is marketed in social network, and a certain piece of information is spread in network. However, in the real world, there are usually many kinds of subjects being spread at the same time in a single social network. In the process of spread, subjects of different kinds may not interacted, but it is more interesting that two or more kinds of subjects may compete with each other, or fight with each other for maximum spread range; or they may help each other and work together to fight for the maximum spread range. In the environment of competition, Influence Maximization problem has more application occasions, and thus attracts more attention and researches.

According to different problem targets, Influence Maximization problems in a competitive environment are basically divided into two kinds: one is to maximize the Influence Spread range of its own; the other is to minimize the Influence Spread range of competitors. These two targets seem to be consistent and complementary, but are substantively different. When the target is to maximize the Influence Spread range of its own, it is not necessary to minimize the spread range of the competitors, or it may even have no effect on the spread range of competitors in extreme cases; conversely, when the target is to minimize the Influence Spread range of competitors, the Influence Spread range of its own may not be maximum. In a simple topological graph of social networks as shown in Figure 12-2, suppose that the spread probability of all edges is 1, two parties involved in the competitive spread are the Red party and the Black party, the Red party selects a node A as a seed node, and the Black party may select two of the remaining nodes as its seed nodes. If the target of the Black party is to maximize its own interests, the Black party will choose the nodes D and E as seed nodes, in this case, the Black party will be able to influence 12 nodes (the Black party realizes the maximum ultimate influence range); if the target of the Black party is to minimize the interests of the competitor, then the Black party will choose the nodes B and C, in this case, the Red party can influence only 6 nodes (the ultimate influence range of the Red party is minimal); if the Black party intends to win in the competition with the Red party, then the Black party will choose the nodes C and D, in this

case, the Red party can influence 9 nodes while the Black party can influence 10 nodes, thereby realizing the Black party's target of winning the competition. Apparently, corresponding seed node selection strategies are needed for different targets of competitive spread.

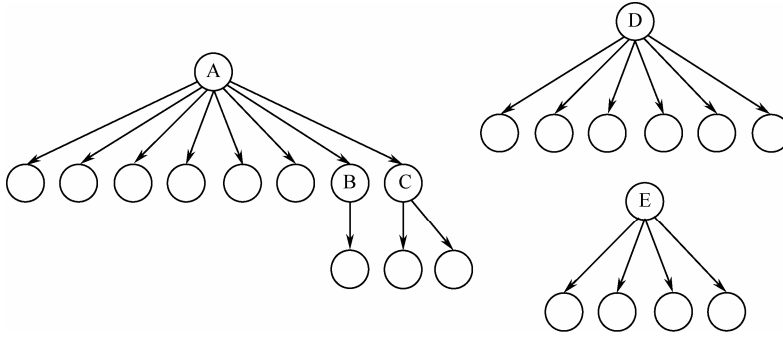


Figure 12-2 Schematic diagram of multi-target competitive spread

The earliest researches on Influence Maximization in competitive environment are from Tim Carnes et al.^[18] in Cornell University and Shishir Bharathi et al.^[19] in the University of Southern California in 2007, and the studying targets of both are to maximize the influence of their own. Bharathi et al. expanded the Independent Cascade Models so as to blend in competition factors, but this author merely provided an approximate algorithm FPTAS based on a special Bi-directed Tree Network Structure. Also in 2007, Carnes et al. presented two types of competitive spread models, i.e. the Distance Model and the Wave Model, and proved that the Competitive Influence Maximization problems in both models were NP-Hard problems, and that an approximate optimal of $1 - 1/e - \varepsilon$ could be realized if the Greedy Hill-Climbing Algorithm can be used. Besides, Wanshou Yang et al.^[20] used the idea of Swarm Intelligence, and adopted Ant Colony Optimization Algorithm to solve the problem of Competitive Influence Maximization. Nam P. Nguyen et al.^[21] had a different studying target. Their studying target is how to identify a minimal number of initial active nodes so as to ultimately make, through diffusion and spread, the influence range of the competitor smaller than a given percentage. The author put forward the basic Hill-Climbing Greedy Algorithm at first, and proved that Greedy Algorithm could reach an approximate optimal result of $1 - 1/e - \varepsilon$. In addition, for the pursuit of higher resolving speed, this author proposed a heuristic strategy based on community, and testified the effectiveness of this method through experiments.

As for the minimization of the influence range of the opponent, Ceren Budak et al.^[22]

from University of California Santa Barbara initiated relevant researches in 2011. On the basis of expanded Independent Cascade Model, Budak et al. proved that the problem of Competitive Influence Minimization is an NP-Hard problem, and compared the performances of Greedy Algorithm with three Heuristic Algorithms. Besides, Xinran He et al.^[23] took into consideration multi-topic competitive factor in Linear Threshold Model, and proved that the objective function in the problem of Competitive Influence Minimization in Linear Threshold Model conformed to the sub-modular character, as a result of which an approximate optimal of $1 - 1/e - \varepsilon$ can be realized when the Greedy Hill-Climbing Strategy is used; in order to enhance the calculation efficiency of Greedy Hill-climbing method, Xinran He et al. put forward an efficient heuristic algorithm CLDAG to make up the deficiency of long run time of Greedy Algorithms. In 2012, scholars including Jason Tsai et al.^[24] studied the problem of Competitive Influence Minimization on the basis of Game Theory, and designed a heuristic mixed strategy for problem solution.

12.7.2 Deformation of Influence Maximization

Traditional problem of Influence Maximization tries to identify K initial active nodes so as to reach the maximum influence range; however, Amit Goyal et al. studied how to choose minimum initial active nodes so as to ultimately influence a given number of nodes. Amit Goyal et al.^[25] proved that the objective function of this problem conformed to the sub-modular characteristics, as a result of which Greedy Hill-Climbing Algorithm can be used for approximating the solution. Moreover, Cheng Long et al.^[26] proved that this problem was an NP-Hard problem. In addition, the author also considered the situation when the target node set is the overall network, and provided corresponding solving algorithm.

Furthermore, Amit Goyal et al.^[25] from the University of British Columbia also studied the minimization problem of Influence Spread time, namely, in order to ultimately influence a given number of nodes, how to identify k nodes so as to activate a given number of nodes in social networks in a shortest time. Similarly, the author provided basic Greedy Algorithms to solve this problem.

Besides, Theodoros Lappas et al.^[27] raised a new problem of K-Effector, namely, if a certain node set S in the network $G(V, E)$ is known, how to select an initial active node set with a size of K so that the ultimate active node set is most consistent with the set S after Influence Spread. The author proved that this problem is an NP-Hard problem in

Independent Cascade Models, and provided, based on a specific tree structure, a dynamic programming algorithm to solve the K-Effector problem.

Based on the problem of Influence Maximization, Theodoros Lappas et al.^[28] studied the problem of Team Formation, namely, if a task T (which needs to be finished by different skill sets), an alternative talent set X (each person has his own skill reserves), and a social network of talent set (the weight of edges between person and person represents the interaction price between them; the smaller the price, the more effectively they can cooperate) are given, how to organize teams in the set X and how to find the talent set X' to execute the task T so that the sum of interaction price in the set X' is the smallest. Based on the diagram diameter and the minimal spanning tree, the author defined two different methods of determining the interaction price, and proved that the problems of Team Formation in both methods were NP-Hard problems, and designed corresponding solving algorithms specific to these two methods.

12.8 Summary

For the past few years, with the rapid development of internet and Web 2.0 technology, and as a communication bridge in real human world, social network has become an important media and platform for inter-communication, knowledge sharing and information spread. The problem of Influence Maximization aiming at discovering the most influential node set in social networks, is a key problem in the field of social network analysis, and is widely applied to marketing, advertisement publication, early warning of public sentiment and many other important occasions, so it is of high researching and application values.

The problem of Influence Maximization in social networks and main studying methods are summarized in this chapter. With respect to algorithms, the current work focuses on the Greedy Algorithms and Heuristic strategies. Existing researches have gained some research results in the efficient handling of Influence Maximization problem; however, due to the large scale of social networks, complex connection between nodes, and dynamic variation of the network structure, many new challenges have been brought to the efficient solution of Influence Maximization. Therefore, there are many problems in this field to be further studied.

(1) Currently, there are numerous parallel computation frameworks which have already been widely applied to such field as massive scientific calculation and so on. MapReduce computation framework is good in programmability, has the advantages of

automatic parallelization, load balancing or the like, and can be operated on large-scale clusters, so its computing power is very remarkable. Therefore, the parallel solving algorithm of Influence Maximization based on MapReduce framework is a feasible and meaningful research discovery. The key problem to be solved in this research direction is how to rationally assign the task of computing the Influence Value of each node in social networks to the computing nodes in the computing cluster, so as to ensure few interactive information between nodes and a short time of dependence and waiting.

(2) In the current Independent Cascade Models or Linear Threshold Models of Influence Maximization, the determination of influence probability or influence weight between nodes is on the basis of constant value assumptions, e.g., with an influence probability of 0.01 or an influence weight of 0.1. These assumptions facilitate the modeling and solving of Influence Maximization problem. However, these assumptions are unreasonable in reality. The determination of influence probability and influence weight between users are closely related to the relationship between users, as well as the communication content, users' interest, users' majors and other contents. Therefore, it is an important research subject as for how to rationally provide the influence probability and the influence weight in the Influence Spread Models. At present, there is a lack of in-depth research of this problem, so this problem may become next research object of interested readers.

References

- [1] David Kempe, Jon Kleinberg, Eva Tardos. Maximizing the spread of influence through a social network [C]. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003: 137-146.
- [2] Jacob Goldenberg, Barak Libai, Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth[J]. Marketing letters, 2001, 12(3): 211-223.
- [3] Mark Granovetter. Threshold models of collective behavior[J]. American journal of sociology, 1978, 83(6): 1420.
- [4] Wei Chen, Yajun Wang, Siyu Yang. Efficient Influence Maximization in social networks [C]. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009: 199-208.
- [5] Eyal Even-Dar, Asaf Shapira. A note on maximizing the spread of influence in social networks [J]. Internet and Network Economics, 2007: 281-286.

- [6] Gerard Cornuejols, Marshall L. Fisher, George L. Nemhauser. Exceptional Paper—Location of Bank Accounts to Optimize Float: An Analytic Study of Exact and Approximate Algorithms [J]. *Management Science*. 1977, 23 (8): 789-810.
- [7] George L. Nemhauser, Lawrence A Wolsey, Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I [J]. *Mathematical Programming*, 1978, 14 (1):265-294.
- [8] Pedro Domingos, Matt Richardson. Mining the network value of customers [C]. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001: 57-66.
- [9] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, Natalie Glance. Cost-effective outbreak detection in networks [C]. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007: 420-429.
- [10] Yu Wang, Gao Cong, Guojie Song, Kunqing Xie. Community-based Greedy Algorithm for mining top-k influential nodes in mobile social networks [C]. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010: 1039-1048.
- [11] Amit Goyal, Wei Lu, Laks V.S. Lakshmanan. Celf++: optimizing the Greedy Algorithm for Influence Maximization in social networks [C]. In *Proceedings of the 20th international conference companion on World wide web*, 2011: 47-48.
- [12] Liu Xiaodong, Li Mo, Li Shanshan, Peng Shaoliang, Liao Xiangke, Lu Xiaopei. IMGPU: GPU Accelerated Influence Maximization in Large-scale Social Networks [J]. *IEEE Transactions on Parallel and Distributed Systems*, 2014: 1.
- [13] Wei Chen, Chi Wang, Yajun Wang. Scalable Influence Maximization for prevalent viral marketing in large-scale social networks [C]. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010: 1029-1038.
- [14] Wei Chen, Yifei Yuan, Li Zhang. Scalable Influence Maximization in social networks under the Linear Threshold Model [C]. In *Proceedings of IEEE 10th International Conference on Data Mining*, 2010: 88-97.
- [15] Amit Goyal, Wei Lu, Laks V. S. Lakshmanan. Simpath: An efficient algorithm for Influence Maximization under the Linear Threshold Model [C]. In *Proceedings of IEEE 11th International Conference on Data Mining*, 2011: 211-220.
- [16] Kyomin Jung, Wooram Heo, Wei Chen. IRIE: A Scalable Influence Maximization Algorithm for Independent Cascade Model and Its Extensions [J]. *arXiv preprint arXiv:1111.4795*, 2011.
- [17] Liu Xiaodong, Li Shanshan, Liao Xiangke, Peng Shaoliang, Wang Lei, Kong Zhiyin. Know by a Handful the Whole Sack: Efficient Sampling for Top-K Influential User Identification in Large Graphs, *World Wide Web Journal*.

- [18] Tim Carnes, Chandrashekhar Nagarajan, Stefan M. Wild, Anke van Zuylen. Maximizing influence in a competitive social network: a follower's perspective [C]. In Proceedings of the ninth international conference on Electronic commerce, 2007: 351-360.
- [19] Shishir Bharathi, David Kempe, Mahyar Salek. Competitive Influence Maximization in social networks [J]. Internet and Network Economics, 2007: 306-311.
- [20] Wan-Shiou Yang, Shi-Xin Weng. Application of the ant colony optimization algorithm to competitive viral marketing [C]. In Proceedings of the 7th Hellenic conference on Artificial Intelligence: theories and applications, 2012: 1-8.
- [21] Nam P. Nguyen, Guanhua Yan, My T. Thai, Stephan Eidenbenz. Containment of misinformation spread in online social networks [J]. Proceedings of the 4th ACM Web Science (WebSci'12), 2012.
- [22] Ceren Budak, Divyakant Agrawal, Amr El Abbadi. Limiting the spread of misinformation in social networks [C]. In Proceedings of the 20th international conference on World wide web, 2011: 665-674.
- [23] Xinran He, Guojie Song, Wei Chen, Qingye Jiang. Influence blocking maximization in social networks under the competitive Linear Threshold Model technical report [J]. arXiv preprint arXiv:1110.4723, 2011.
- [24] Jason Tsai, Thanh H. Nguyen, Milind Tambe. Security games for controlling contagion [C]. In Proceedings of the Twenty-Sixth National Conference in Artificial Intelligence, 2012.
- [25] Amit Goyal, Francesco Bonchi, Laks V. S. Lakshmanan, Suresh Venkatasubramanian. On minimizing budget and time in influence propagation over social networks [J]. Social Network Analysis and Mining, 2012: 1-14.
- [26] Cheng Long, Raymond Chi-Wing Wong. Minimizing Seed Set for Viral Marketing [C]. In Proceedings of the IEEE 11th International Conference on Data Mining, 2011: 427-436.
- [27] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, Heikki Mannila. Finding effectors in social networks [C]. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010: 1059-1068.
- [28] Theodoros Lappas, Kun Liu, Evimaria Terzi. Finding a team of experts in social networks [C]. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009: 467-476.